„Hadoop & IT-Strategie –

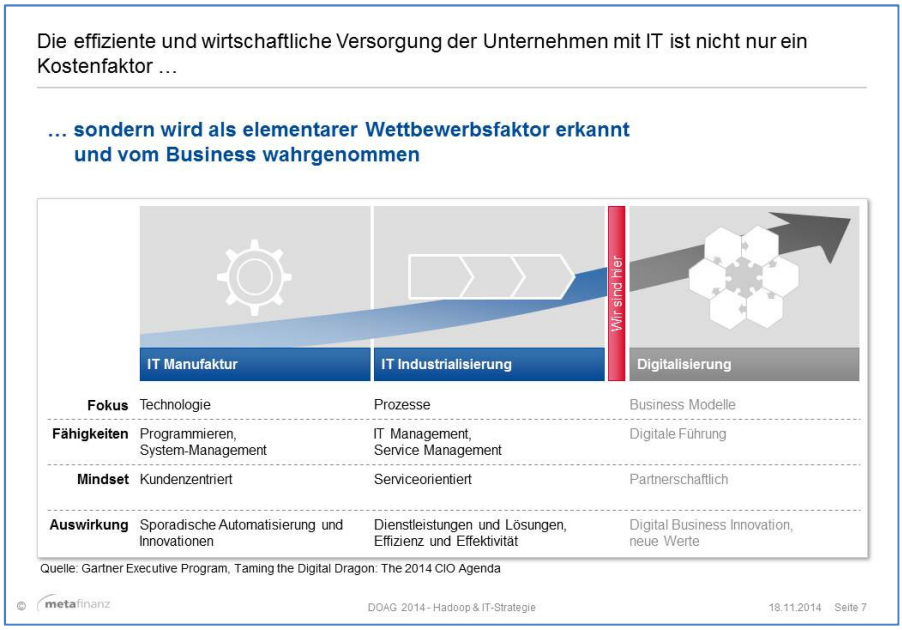
Ein Spagat zwischen Innovation und Kosten – geht das überhaupt?

Oliver Herzberg
Metafinanz-Informationssysteme GmbH
München

Schlüsselworte
IT-Strategie, Digitalisierung, Innovation, Kostendruck, regulatorische Anforderungen, SolvencyII, IFRS, Hadoop, Datawarehouse, Vorgehensmodell, Enterprise Architecture, DevOps

Standortbestimmung & Status Quo

Die Unternehmens-IT steht an einer Evolutionsschwelle. Die Individual- und Maßfertigung isolierter und unternehmensspezifischer Lösungen rückte in den letzten Jahren vermehrt in den Hintergrund, viele Unternehmen stellen die Industrialisierung ihrer IT in den Mittelpunkt.



Quelle: Gartner Executive Program, Taming the Digital Dragon: The 2014 CIO Agenda

Die Ausrichtung der IT wurde deutlich serviceorientierter, interne Prozesse sowie das Zusammenspiel mit dem Business wurden neu organisiert. ITIL als der Standard für IT-Prozesse fand und findet in vielen Unternehmen Einzug.

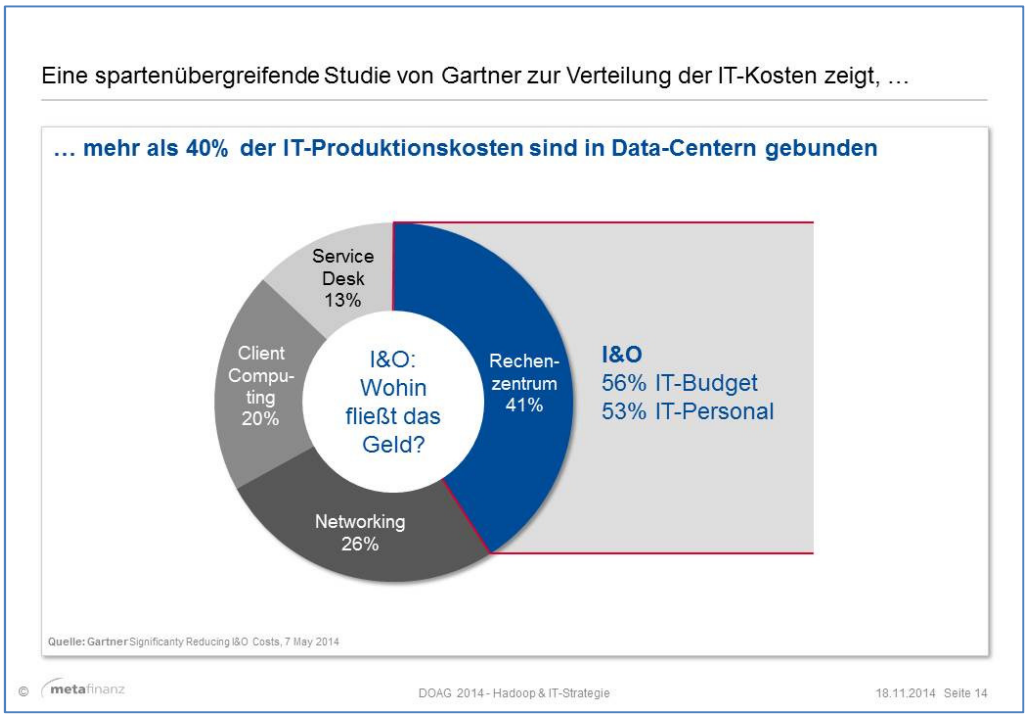
D.h.

- die Einführung standardisierter und arbeitsteiliger Prozesse,
- die Aufstellung der IT-Organisation entlang der Wertschöpfungskette,
- aufgeteilt in Run und Change, teilweise unter Einbeziehung (lokationsübergreifender) shared Service Einheiten,
- eingebunden in eine unternehmensweite Sourcingstrategie (Lokationen, Lieferanten, Leistungen),
- sowie die professionelle Aufstellung Richtung Kunde

stehen im Mittelpunkt.

Dennoch gehören, gerade bei Finanzdienstleistern die IT-Kosten weiterhin zu den wesentlichen Kostenblöcken im Unternehmen. Die IT-Verantwortlichen stehen damit vor einer großen Herausforderung, insbesondere auch vor dem Hintergrund der globalen gesamtwirtschaftlichen Lage.

Die effiziente und wirtschaftliche Versorgung der Unternehmen mit IT wird aber nicht nur als Kostenfaktor sondern als elementarer Wettbewerbsfaktor erkannt und vom Business wahrgenommen.



Quelle: Gartner Significanty Reducing I&O Costs, 7 May 2014

Eine spartenübergreifende Studie von Gartner zur Verteilung der IT-Kosten zeigt, dass mehr als 40% der IT-Produktionskosten in den Data Centern gebunden sind. Danach folgt Netzwerkinfrastruktur und der Bereich der Clients.

Neben der Umsetzung kurzfristig wirksamer Kostensenkungsmaßnahmen (z.B. Provider-/Einkaufsmanagement) gilt es hier, die Kostenstrukturen auch nachhaltig zu optimieren.

Unsere Erfahrungen aus der Finanzindustrie bestätigen dieses Ergebnis. So investiert ein großer globaler Versicherungskonzern derzeit in die weltweite Konsolidierung seiner Datacenter. Damit geht die Erneuerung der Netzwerkinfrastruktur einher, sowie der Austausch der Arbeitsplätze durch den weltweiten Rollout virtueller Clients.

Weitere Stellschrauben für ein erfolgreiches und nachhaltig wirksames Kostenmanagement sind die Standardisierung der IT, die Etablierung von Steuerungsmodellen und ein leistungsfähiges Portfoliomanagement.

Auf der **Infrastrukturseite** laufen viele Investitionen in die Modernisierung historisch gewachsener und teilweise überalterter IT-Anwendungen. Zahlreiche Finanzdienstleister investieren in die Standardisierung, Erneuerung und Konsolidierung ihrer IT-Kernsysteme z.B. Kreditsysteme im Bankenbereich sowie in die Erneuerung ihrer in die Jahre gekommener Datawarehouses.

Die Einführung leistungsfähiger, flexibler und elastischer Datawarehouses und BI-Infrastrukturen wirkt vor allem in Changekosten. Unsere Erfahrungen zeigen, dass bei großen Projekten-/Programmen bis zu 60% der Aufwände direkt oder indirekt der Datenbeschaffung- /validierung und /-transformation zugeordnet werden können. Dazu kommen komplexe und aufwändige Testphasen.

Stetig wachsende Datenmengen z.B. aufgrund Solvency II, CRD IV, IFRS, EBA-Reporting, Stresstesting, Adjustierung interner Modelle und Compliance wirken direkt auf den Speicherbedarf und die damit in Zusammenhang stehenden IT-Kosten.

Die kontinuierlich steigenden **regulatorischen Anforderungen** haben wesentliche Auswirkungen auf das Geschäft von Finanzdienstleistern. Mit Solvency II wird z.B. in der Assekuranz eine höhere Transparenz, ein besseres Risikomanagement, eine Solvenzregelung (mit Frühwarnsignalen und dazugehörigen Instrumenten) geschaffen.

Einer Lünedonk Studie zur Folge wird die Umsetzung der Regulierungsauflagen rund um Solvency II die Versicherungswirtschaft bis 2020 auch weiterhin belasten. Eine Herausforderung stellen die steigenden Datenvolumina dar: Für 2020 erwarten die Versicherer nicht nur eine massive Zunahme des Datenvolumens, sondern auch der Investitionen in die Analyse der Daten (Infrastruktur, Ressourcen, Knowhow), um beispielsweise Kunden- und Risikogruppen besser segmentieren zu können.

Mit IFRS9 wurde im Juli 2014 nach mehrjähriger Entwicklungsphase und mehrfachen Terminverschiebungen eine wesentliche Änderung zur Bilanzierung von Finanzinstrumenten verabschiedet.

Die Implementierung und laufende Erfüllung dieser regulatorischen Anforderungen bindet nicht nur Ressourcen in IT und Fachbereichen, sondern stellt für die Unternehmen vor allem eine hohe Kostenbelastung dar.

Kosten, die dann im Gesamtbudget der zur Verfügung stehenden Mittel, den Raum für Innovationen und Fortschritt deutlich schrumpfen lassen. Auch deshalb, weil es der externe Umsetzungsdruck oftmals nicht erlaubt, nachhaltige Lösungen zu entwickeln. Der Betrieb und die spätere Ablösung der dann eingesetzten Zwischenlösungen zehren weiter an den verfügbaren Mitteln.

Der Markt für Finanzdienstleister befindet sich im Umbruch. Die Kundenbedürfnisse verändern sich rasch, die Marktanforderungen werden immer komplexer.

Folgende Marktstrukturveränderungen werden untermauert durch eine Lünedonkstudie am Beispiel der Versicherungswirtschaft bis 2020 vorausgesehen: mehr Zusammenschlüsse zwischen Versicherungen als bisher, aber auch Kooperationen mit Banken zur Erschließung des Geschäfts mit dem Asset- und Vermögensmanagement. Stärkerer Einsatz mobiler Technologien über mobile Plattformen in der Kundenkommunikation z.B. durch mobile Apps im Vertrieb und Service von Versicherungsprodukten, dem Direktvertrieb über Online-Portale sowie stringente Multi-Channel-Kommunikation zur ganzheitlichen Kundenansprache über mehrere Kanäle [1].

Daten und Informationen sind also die Grundlage, der Rohstoff für einen effizienten, wirtschaftlichen und vor allem erfolgreichen Geschäftsbetrieb von Finanzdienstleistern.

In Folge all dieser Entwicklungen müssen Unternehmen heute immer größere Anstrengungen unternehmen, um explodierenden Datenmengen Herr zu werden.

[1] Quelle: Lünedonk, Trendstudie Versicherungen

Durch **die Digitalisierung** ganzer Lebensbereiche wird eine bislang unbekannte Datentiefe ermöglicht. Die Vielfalt, Herkunft, Verfügbarkeit, Qualität und letztlich die gewonnen Aussagekraft variieren signifikant.

Die Ausbreitung digitaler Wertschöpfungsaktivitäten und insbesondere moderner Informations- und Kommunikationstechnologien hat in den letzten Jahren enorm an Dynamik gewonnen.

Deutschland hat bei der Digitalisierung der Wirtschaft in fast allen Branchen Nachholbedarf. Zu diesem Ergebnis gelangt eine Studie der Boston Consulting Group für das manager magazin (18.Juli 2014). Darin verglichen die Strategieberater sieben deutsche Kernindustriezweige, die etwa 80 Prozent der Dax-Konzerne und 70 Prozent der M-Dax-Unternehmen abdecken, mit dem jeweiligen digitalen Branchenvorbild.[2]

Führende Politiker rufen dazu auf, die Digitalisierung mit größerer Entschlossenheit vorzutreiben. „Wir müssen die Geschwindigkeit unseres Handelns deutlich erhöhen. Die Revolution vollzieht sich schneller, als es viele Akteure in Politik und Wirtschaft wahrhaben wollen“ sagt der designierte EU-Kommissar für digitale Wirtschaft, Günther Oettinger der „Welt am Sonntag“ (14.09.2014).

Unternehmen wird damit eine Zukunftsaufgabe gestellt: Sie müssen dem strategischen Management der internen und externen Daten und Informationen als auch deren Umsetzung in der Praxis einen höheren Stellenwert zumessen.

Gleichbedeutend ist die Organisation von betrieblichen Strukturen, die den Umgang mit **Sicherheit und der Vertraulichkeit von Daten** und Informationen gewährleisten. Damit sind zum einen unternehmensinterne Mechanismen wie z.B. die Einführung von Security Policys gemeint. Dort werden u.a. der Schutzbedarf der Daten und Informationen sowie der damit verbundenen Maßnahmen geregelt. Zum anderen müssen Mechanismen entwickelt werden, die den externen Zugriff auf diese Daten steuern und die Informationen schützen.

Datenmanagement wird damit elementaren Bestandteil der IT-Strategie von Unternehmen.

CIOs stehen damit vor der Herausforderung,

- der Regulatorik genüge zu leisten, was Kosten ohne Ertrag bedeutet
- die Chancen aus der Digitalisierung zur Bindung und Erschließung neuer Märkte gewinnbringend zu nutzen
- andererseits aber auch aktiv die Potentiale zu erkennen und zu nutzen, die aus solchen Transformationsprozessen entstehen: Gewaltige Datenschätze, die es zu heben gilt.

Und hier liegt das Problem: Die gewaltigen Daten werden erst dann zu Schätzen, wenn man weiß, wie sie für das eigene Unternehmen gewinnbringend genutzt werden können. Keine leichte Aufgabe.

Eine besondere Herausforderung kommt der Frage zu, welchen Trends es zu folgen gilt und welche Trends Modeerscheinungen ohne konkrete Wertschöpfung bleiben.

[2] Quelle: manager magazin 07/2014

Entscheider aus IT und Fachbereichen stehen vor der Herausforderung, die IT Produktionskosten nachhaltig senken zu müssen und gleichzeitig dem Business durch innovative Lösungen Wettbewerbsvorteile ggü. Mitbewerbern zu verschaffen.

Doch wie kann der Spagat zwischen Innovation und Kosten dennoch nachhaltig gelingen?

Vor dem Hintergrund der dargestellten Rahmenbedingungen für Unternehmen, soll die Tauglichkeit von Hadoop-Lösungen näher beleuchtet werden.

Mit dem Opensource Framework Hadoop stehen neue und bisher nicht umsetzbare Analytics Möglichkeiten zur Verfügung. Es gibt aber auch konkrete Usecases, die auf die Reduzierung von Infrastruktur-/Storagekosten abzielen.

Hadoop besteht im Kern aus zwei Komponenten: einem verteilten Filesystem (HDFS), welches auf einem Cluster aus Standard-Servern liegt und beliebige Daten in beliebigen Formaten speichern kann, sowie dem MapReduce-Framework zur parallelen Verarbeitung dieser Daten.

Rund um diesen Kern ist ein ganzes Ökosystem von Tools entstanden. Manche von ihnen bieten die Möglichkeit, Daten mit SQL auszuwerten. In diesem Zusammenhang sei auf den unsere Veröffentlichung zum Thema „Datenaustausch Hadoop & Oracle DB“ www.metafinanz.de/sites/default/files/DOAG_2013_Datenaustausch%20Hadoop%20%26%20Oracle%20DB.pdf verwiesen.

Die Stärken von Hadoop sind die kostengünstige Speicherung beliebig strukturierter und nicht strukturierter Daten sowie die parallele Verarbeitung von riesigen Datenmengen. Hadoop Cluster skalieren linear in Bezug auf Speicherkapazität und Performance. 10% mehr Knoten bringen also 10% mehr Speicherkapazität und auch 10% mehr Performance.

Think big, start small – gilt auch für Hadoop

Hadoop – Lösungen können in der Praxis in drei Gruppen eingeteilt werden:

RDBMS Offload

Hierunter fallen Projekte die ihren Fokus auf die Optimierung bereits bestehender Enterprise Data Warehouse Systeme legen. Insbesondere wird hier das Augenmerk auf die Verkürzung bestehender ETL Strecken gelegt, was durch eine Auslagerung in einen Hadoop Cluster erreicht wird. Als Folge wird das Data Warehouse entlastet, Lizenz- wie auch Speicherkosten sinken und DWH basierende Anwendungen erreichen bedeutend mehr Performance ohne weiteres Investment in die existierende DWH Infrastruktur. In allen Fällen ist der ROI noch während des ersten Geschäftszyklus erreicht.

DHW Extension

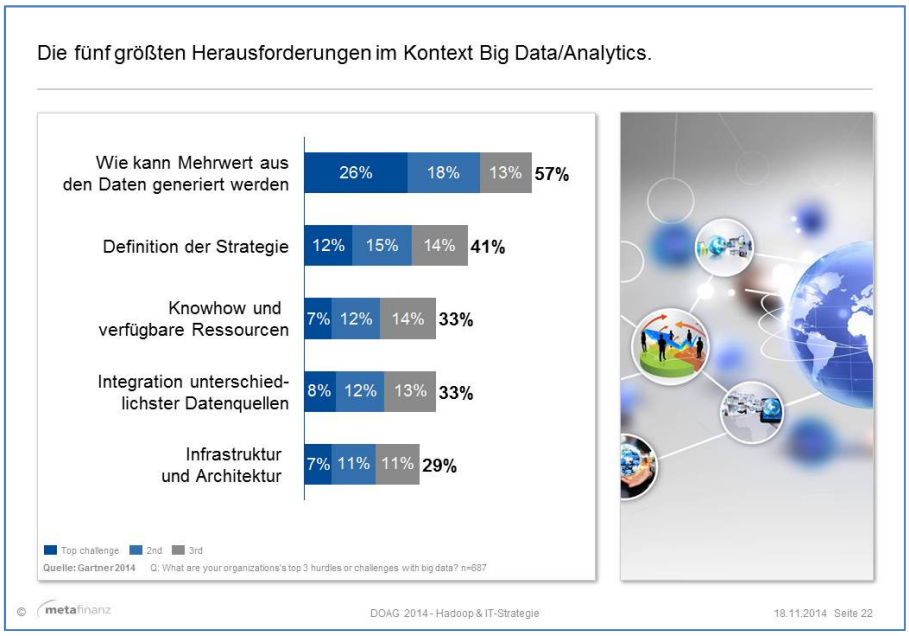
Hier werden bestehende Datawarehouse Lösungen durch den Einsatz von Hadoop funktional erweitert. Dies erfolgt z.B. dadurch, dass Operational Data Stores (ODS) als vorgeschalteter „Datensammler“ durch Hadoop ersetzt werden.

Big Data Exploration

Die Sammlung von Daten aus unterschiedlichen Quellen und die Verknüpfung mit unterschiedlichen Datentypen ordnen wir dem Typ „Big Data Exploration“ zu. Auch der Einsatz von zum Beispiel Data Mining und Machine Learning fallen darunter.

Ungeachtet der Klassifizierung, sieht ein Großteil der Unternehmen einer Gartner Studie zur Folge die größte Herausforderung im Kontext Big Data darin, welcher Mehrwert aus den Daten generiert

werden kann, gefolgt von der Frage was Big Data eigentlich ist und wie sie von dieser Technologie partizipieren können, ohne eine Investmentkollaps zu erleiden. .



Quelle: Gartner Executive Program, Taming the Digital Dragon: The 2014 CIO Agenda

Hier wird das Dilemma deutlich. Zwar sehen aus einer gewissen Flughöhe Hadoop-Projekte wie klassische DWH/BI-Projekte aus, wenn man jedoch tiefer geht, ergeben sich Unterschiede zur klassischen BI.

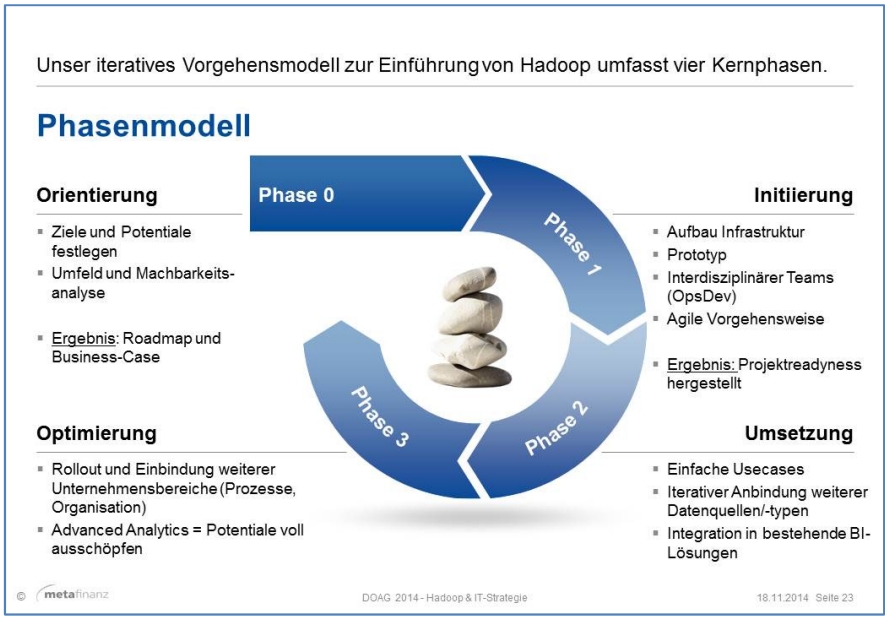
In der klassischen BI-Welt können die Projektziele z.B. der Aufbau eines Neartime-Reportings zur Aktiv-/Passivsteuerung, ein Management-Informationscockpit oder eine Dashboard zum GuV-Stresstesting relativ klar definiert werden. Das Business hat Fragestellungen zu beantworten, die IT strukturiert diese Daten und liefert Lösungen dafür.

Bei Hadoop, insbesondere „Data Exploration“ Projekten sieht es ein wenig anders aus. Hier stellt die IT oftmals eine „Forschungs- und Entwicklungsplattform“ zur Verfügung, auf der das Business herausfinden kann, welche Fragen gestellt werden können.

Das hört sich vielleicht ein wenig abgehoben an, wenn wir uns damit näher beschäftigen wird klarer was damit gemeint ist.

Unser in der Praxis bewährtes iteratives Vorgehensmodell zur Einführung von Hadoop umfasst vier Kernphasen:

Phasenmodell



Orientierungsphase

In dieser Phase liegt der Schwerpunkt darin eine grundsätzliche Strategie für den Einsatz von Hadoop im Unternehmen zu erarbeiten. Eine Heatmap über mögliche Business Potentiale zu erstellen und dazugehörige Beispielusecases zur Validierung in Folgenphasen zu entwickeln.

Wesentlicher Aspekt ist die Stakeholder zu identifizieren und von der Investition in das Ungewisse zu überzeugen:

- Erarbeitung der Ziele und Potentiale von Big Data im Unternehmen ausgehend von den geschäftlichen Anforderungen/Nutzen
- Umfeld- und Machbarkeitsanalyse (Standortbestimmung)
- Daten, Systeme, Infrastruktur, Organisation, Sicherheit, vorhandene Skills und notwendige Knowhow, ggf. über Partner
- Zielbild und Lösungsdesign
    - Roadmap (Prozesse, Daten, Technologien, Kompetenzen)
    - Erarbeitung erster Use Cases
    - Betriebs-/Wartungskonzept, Kosten-/Nutzenbetrachtung, Nutzungskonzept
    - Vorüberlegungen zu künftigen Betriebs-/Sourcing-/Lizenzmodellen
    - Sicherheitsstrategien und Konzepte
    - Notwendige Kompetenzen

Initierungsphase

Nach dem sich das Unternehmen für den Einsatz von Hadoop entschieden hat, geht es nun darum, schnell die Project-Readyness herzustellen. Hierzu gehören die technische Funktionsfähigkeit sowie die richtige Teamzusammensetzung. Ansätze aus der DevOps – Bewegung, bei der interdisziplinär das Knowhow aus den unterschiedlichsten Domänen wie Business, Entwicklung und Betrieb gebündelt werden, tragen entscheidend zum Projekterfolg bei.

- Aufbau Infrastruktur
  - Hardware (Netzwerk, Server), Software, Zugang zu externen Datenquellen
  - Infrastruktur zur Datenspeicherung, Identifizierung von Tools und Frameworks, die es sinnvoll und bedarfsweise zu integrieren gilt
- IT-Organisation
  - Auswirkungen IT Betrieb – Rollen- und Servicemodell /-management
- Bildung interdisziplinärer Teams (Data Scientist, Business-/Datenanalysten, Entwickler, Nutzer, Entscheider)
- Schulung & Skillaufbau
- Klare Verantwortlichkeiten festlegen (Business/IT)
- Datenschutz und Securityaspekte konkretisieren
- Erste Gehversuche (Tests, Skills, Technologien), Setup einfacher und kostengünstiger Use Cases, vertraut werden mit „Technologie, Tools, Daten“ z.B. durch
  - Upload klassischer RDBM-Systeme und Analyse in Hadoop
  - Logfiles für Security – Auswertungen
  - Archivierung und Validierung von Daten unterschiedlicher, aber bekannter Quellen

Wir empfehlen insbesondere die Bildung von interdisziplinären Teams aus IT, Betriebsorganisation und Business.

Umsetzungsphase(n)

Die Umsetzung der Usecase erfolgt iterativ, agil und inkrementell. Als Methodik hat sich hier Scrum mit regelmaessigen Sprints bewährt. Hadoop wird schrittweise in die Organisation (Daten, Prozess, Infrastruktur, Management) eingeführt, die analytischen Fähigkeiten sowie die Datenqualität damit sukzessive aufgebaut und erweitert.

Iteration 1 – Quick Wins

- Beginn der Implementierung des ersten Use Cases zur Erzielung von Quickwins
- Arbeiten mit bekannten Datenquellen und Informationen
- Neue Erkenntnisse finden und validieren/interpretieren

Iteration 2 – n

- Ausweitung der Use Cases (Datenmenge, -quellen, -volumen, externe Daten)
- Datenextraktion und Aufbereitung (ETL), denn auch im Hadoop Umfeld liegt dort der größte Teil des Aufwands
- Erarbeitung IT-Service-/Betriebsmodell und Anpassung bestehenden IT-Prozesse
- Integration der Hadoop Lösung in bestehende BI-Lösungen
- Advanced Analytics (Musterkennung, Statistik, Datamining, Forecast – arbeiten mit den Daten)
- Stabilisierung der Wertschöpfungsprozesse (Datenbewirtschaftung, Sicherheit, Performance, …)
- Fitmachen der Organisation für den Einsatz von Hadoop

Optimierungsphase

In der Optimierungsphase gilt es den Betrieb zu stabilisieren und den Nutzen von Hadoop breit zu realisieren:

- Rollout und Einbindung weiterer Unternehmensbereiche
- Ausbau Knowhow
- Unternehmensweite Verfügbarkeit
- Ergebnisse nutzbar machen
- Hadoop Potentiale voll ausschöpfen und wirtschaftlichen Nutzen realisieren

An folgenden Anwendungsszenarien soll veranschaulicht werden, wie durch die Einführung einer Hadoop-Lösung nachhaltige, aber auch gleichzeitig innovative Assets geschaffen werden können.

Szenario 1: Proof of Concept "strategische Personalentwicklung" Zielsetzung: Quickwins realisieren

Ausgangssituation

Im Rahmen der gezielten Personalsuche muss schnell und passend auf Kundenanfragen reagiert werden. Hier ist zu berücksichtigen, ob die angefragten Skills im Unternehmen grundsätzlich vorhanden und für den Einsatz beim Kunden verfügbar sind.

Im Rahmen der strategischen Personalentwicklung müssen die Skills der Mitarbeiter permanent an den Marktbedürfnissen ausgerichtet werden. Das betrifft zum einen die Weiterentwicklung des bestehenden Personalstamms als auch des gezielten Recruitings (sowie Werben) von Nachwuchskräften.

Herausforderung

Die zur Verfügung stehenden Informationen, wie z.B. CV's (Lebensläufe), Projekthistorien, Schulungspläne, Bedarfe aus dem Markt, Gehaltsspiegel der Branche, Stundensätze, interne und externe Stellenausschreibungen, liegen strukturiert (z.B. HR-Stammdaten) als auch unstrukturiert (z.B. Stellenausschreibungen) vor. Teilweise handelt es sich um personenbezogene Daten, deren besondere Klassifizierung spezielle Sicherheitsverfahren erfordern.

Zu Projektbeginn ist unklar, in wie weit sich die Daten überhaupt in Beziehung setzen lassen und miteinander korrelieren.

Lösungsansatz mit Hadoop

In einem ersten Schritt wird ein Prototyp auf Hadoop-Basis in einer definieren Projektumgebung entwickelt. Bei der Entwicklung wird auf den Einsatz von OpenSource Technologien geachtet, um ein Vendor-LockIn zu vermeiden. Die Mitarbeiterprofile werden nach HDFS geladen und dort mit MapReduce in verteilte Index(e) gesplittet. Die Auswertungen werden über Velocity, Solr Console, REST API und SorjJ auf den Originaldaten ermöglicht. Durch den Einsatz von SolR Cloud erreicht der Kunde ein hohes Maß an Flexibilität und Systemstabilität, was sich in einer recht geringen TCO widerspiegelt.

Bereits jetzt ist das Unternehmen in der Lage bei Kundenanfragen schnell feststellen zu können, ob das geforderte Knowhow grundsätzlich im Unternehmen verfügbar ist.

In einem zweiten Schritt werden nun die Daten aus dem ERP-System integriert. D.h. die Frage in welchem Projekt der Mitarbeiter aktuell mit welcher Laufzeit engagiert ist.

In einem dritten Schritt werden extern verfügbar Daten aus öffentlich zugänglichen Stellenbörsen und OpenData Portalen integriert. D.h. es kann über das breite Hadoop-Spektrum mit seinen vielfältigen Analysetools analysiert werden, ob es sich um Spezialwissen oder am Markt breit verfügbares Wissen handelt.

Der Kunde ist jetzt zum einen in der Lage schnell auf angefragte Ressourcen reagieren zu können, zum anderen kann er seine Skills gezielt strategisch weiter entwickeln und an den Bedürfnissen seiner Kunden auszurichten.

Szenario 2: Zentralisierung Archivierungsmedien und Reduzierung Archivierungskosten; Zielsetzung: Speicherkosten reduzieren und Daten auswertbar machen

Ausgangsituation

Zur Erfüllung u.a. der GOBS (Grundsätze ordnungsgemäßer Buchführungssysteme) oder der GdPdU (Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen) müssen Geschäftsrelevante Daten 10 Jahre archiviert werden.

Herausforderung

Die Archivierung der Daten erfolgt oft im Datawarehouse oder in speziellen Archivierungsmarts durch den Einsatz von RDBMS und klassischen Speichermedien (SAN, Band, optische Medien, ..). oder auch in klassischen Archivierungssystemen.

Die Speicherung ist teuer und umstäendlich. Um Kosten zu reduzieren werden die Daten oft verdichtet, mit der Folge das Langzeitdaten meistens für eine Analyse unbrauchbar geworden sind und nur noch gesetzlichen Anforderungen genügen. Es geht wertvolles Wissen über Unternehmensentwicklung verloren. Darüber hinaus sind die Daten, die wertvolle Geschäftsinformationen enthalten, im Laufe der Zeit schwer auswertbar, da sich die Basissoftware der Systeme weiterentwickelt, sich DB-Schemata ändern und Systeme im Falle des Zugriffs (z.B. im Prüfungsfall) nur unter hohem Aufwand in der Lage sind die gespeicherten Daten erneut zu laden und zu verarbeiten.

Archivierungssysteme folgen oftmals ihrer eigenen Logik der Datenablage und –indizierung. D.h. zum Zeitpunkt der Archivierung werden die Ordnungsbegriffe, nach denen Daten später gesucht werden, festgelegt. Eine spätere Neuindizierung ist zwar möglich, aber mit entsprechenden Aufwänden verbunden.

Lösungsansatz mit Hadoop

Data Warehouse Architekturen ähneln im Prinzip den Bereichen Quelle, Staging, Core und Data Marts.

Die im Core-DWH in einer RDBMS gehaltenen Daten werden in den Hadoop Cluster repliziert. Die bestehenden ETL-Prozesse zur Befüllung der Data Marts bleiben unverändert.

1. Schritt - Quickwin durch Reduzierung Speicherkosten im Core DWH:

Nach der initialen Befüllung in Hadoop könnte man – je nach Anwendungsszenario – z.B. alle Daten zu archivierenden Daten aus dem Core-DWH in ein Hadoop-Archiv verschieben. Die Datenvorhaltung des Core-DWH wird kleiner, die TCO und die lfd. Speicherkosten sinken. In der Folge sinkt sowohl die Dauer eines Backups als auch die Restorezeit, die meist an SLA's gebunden sind.

Dieses Beispiel wird im morgigen Vortrag Data Mart (Star Schema) Offload nach Hadoop meines Kollegen näher erläutert und weiter detailliert.

2. Schritt - Ausbau zum Datenarchiv und Ablösung klassischer Archivsysteme:

In einem nächsten Schritt werden die im Archivsystem gespeicherten Daten in das Hadoop-Cluster geladen. Nun liegen die Daten unterschiedlichster Formate (pdf, word, excel, xml, …) auf dem Hadoop-Cluster und können durch den Einsatz verschiedenster Komponenten des Hadoop Ökosystems analysiert und verarbeitet werden. Durch eine geschickte Verbindung zwischen Core-DWH und dem Hadoop basierenden Archivierungssystem ist es möglich, innerhalb einer Abfrage sowohl auf die Bestands- wie auch Archivierungsdaten zuzugreifen. Da Hadoop selbst die Daten immuitable anlegt (einmal geschriebene Daten können nur noch gelöscht oder angefügt werden) wird auch gesetzlichen Vorgaben Rechnung getragen.

Szenario 3: Hadoop als Staging-Area im ETL-Prozess – Zielsetzung: Umsetzung gesetzlicher Anforderungen z.B. BCBS 239 bei gleichzeitiger Reduzierung Infrastrukturkosten

Ausgangssituation

Im Staging- und ETL-Prozess eines Finanzdienstleisters werden im Rahmen der Tagesendeverarbeitung große Datenmengen zwischengespeichert. Diese Daten liegen mit in den Geschäftsprozessen begründeten unterschiedlichen Qualitäts- und Reifegraden vor und müssen validiert und ggf. angereichert werden.

Die Belieferung der Daten erfolgt oftmals in Form von Flat-Files, die dann mittels individuell entwickelter ETL-Prozesse in eine relationale Datenbank geladen werden. Nachdem die Vollständigkeit des Geschäftsdatenbestands festgestellt wurde, werden die Daten mittels Unload unterschiedlichsten Nutzern oder Nebenbuchhaltungssystemen (z.B. SAP - Bankanalyzer) zur Befüllung ihrer Datamarts / Rechenkerne zur Verfügung gestellt. Auf dieser Grundlage werden dann u.a. unterschiedlichste fachlich motivierte Auswertungen und Reports erstellt.

In der Praxis zeigt sich jedoch, dass diese Daten- und Informationsinfrastrukturen nicht ausreichen um den Bedarf an schneller, spezifischer und genauer Information abdecken zu können. In den Fachbereichen finden sich deshalb oft individuelle Lösungen, die den Grundsätzen ordnungsgemäßer DV-Systeme nur bedingt entsprechen. Diese Lösungen greifen oftmals direkt auf die operativen Systeme zu, replizieren Daten und produzieren für sich nicht nur wieder neue Informationen – sondern auch daraus resultierende Kosten.

Ein Überleitbarkeit und Erklärbarkeit von Steuerungsinformation für das Management ist damit nur mit entsprechend hohem Ressourceneinsatz möglich.

Herausforderung

Strukturelle Änderungen am Datenmodell z.B. durch die Einführung von regulatorischen Anforderungen, an den operativen Vorsystemen oder fachlich motivierte inhaltliche Änderungen (z.B. Bewertungsmethoden) sind im „Change“ sehr teuer, da die komplette ETL-Strecke konzipiert, geändert und vor allem durch alle Nutzer getestet werden muss.

Lösungsansatz mit Hadoop

Im Zielbild soll ein dem Core-DWH vorgelagerter „Data-Lake“ aufgebaut werden. Der „Data-Lake“ ist ein Speicher-Repository, das zunächst große Mengen an Rohdaten in ihrem ursprünglichen Format speichert.

Beim Load, also bei der Bereitstellung der Basisdaten durch die operativen Vorsysteme, wird jedem Datenelement eine eindeutige Kennung zugewiesen und mit einem Satz erweiterter Metadaten-Tags ergänzt.

Dadurch wird die Konsistenz der Daten im Warehouse gewährleistet, Peer-to-Peer Schnittstellen können eliminiert werden. Dezentrale Anwendungen in den Fachbereichen können mittels klarer Schnittstellen, über ein Berechtigungskonzept gesteuert sowohl auf die Rohdaten aus den operativen Vorsystemen als auch auf die Daten des DWHs und seiner Marts zugreifen und das, ohne dabei die fachlich motivierte und notwendige Flexibilität zu verlieren.

Quickwin durch Nutzung kostengünstigerer Speichermedien
Durch den Einsatz von Hadoop und HDFS lassen sich damit die Kosten der auf unterschiedlichsten Fileservern und SAN's abgelegten Daten drastisch reduzieren.

Der Vorteil liegt darin, dass relativ günstige Speichermedien des Hadoop-Clusters genutzt werden können, statt der üblicherweise vom RZ-Provider vorgesehenen weit teureren Lösungen.

Aus Analyticssicht entstehen die Vorteile dadurch, dass die Daten nun durch den Einsatz von Hadoop – Technologien genau in der Granularität verfügbar sind, wie sie aus Sicht der fachlichen Nutzer benötigt werden. Dezentrale Anwendungen lassen sich so reduzieren.

Ein großer Schritt Richtung Überleitbarkeit und Transparenz der Berichterstattung, wie sich nicht nur z.B. durch BCBS239 gefordert wird!

Für weiterführende Informationen, möchten ich auf den DOAG Vortrag „**Data Mart Offload nach Hadoop**" verweisen.

Fazit und take away

Zusammenfassend lässt sich feststellen, dass wir, um mit Hadoop erfolgreich zu sein, bereit sein müssen neue Wege durch unbekanntes Terrain zu beschreiten.

Best Practises müssen sich erst noch etablieren bzw. dem jeweiligen Unternehmenszweck angepasst werden.

Wir sind der Meinung, dass für den Einsatz von Big Data Technologien das gleiche gilt, wie für den Einsatz anderer innovativer, neuartiger und damit unbekannter Technologien:

Klein anfangen, ein „Gefühl" für die Technologie und ihre Möglichkeiten entwickeln. Auf diesem Fundament kann eine tragende, an dem jeweiligen Unternehmen und dessen Zweck ausgerichtete Hadoop-Strategie erarbeitet werden.

Datawarehouses als auch relationale Datenbanken werden durch den Einsatz von Hadoop nicht überflüssig, sondern sinnvoll ergänzt. Die Investition in bestehende DWH-/BI- und Reporting-infrastrukturen wird gesichert, das Potential für erweiterte Analyticsmöglichkeiten auf der einen und Cost Savings auf der anderen Seite werden eröffnet, der Weg in die digitale Zukunft geebnet.

Kontaktadresse:
Oliver Herzberg
Metafinanz Informationssysteme GmbH
Leopoldstrasse 146
D-80804 München

Telefon: +49 (0) 89-360 531 5432
Fax: +49 (0) 89-360 531 115
E-Mail Oliver.Herzberg@metafinanz.de
Internet: www.metaffinanz.de