

BI-Architekturen der Zukunft

■ Teil I: (Near-)Realtime und Analytic Excellence

von Rita Hennig und Mathis Manz



Risikoreports erhalten Sie täglich im Batchbetrieb. Reicht Ihnen das? Denn: Die Zukunft sieht anders aus, die Welt verändert sich im Sekundentakt: Hier übernehmen Terroristen die Kontrolle über ein Ölfeld, dort wird Land X im Rating abgestuft und wie lange hält die Immobilienblase noch? Glücklicherweise, wer dabei den Überblick behält.

Drei Schlagwörter bestimmen diesbezüglich den aktuellen Diskurs: (Near-)Realtime, Analytic Excellence und Selfservice-BI, auch bekannt als Sandboxing. Wir haben uns vor dem Hintergrund dieser Schlüsselbegriffe gefragt: Wie sieht die Business-Intelligence(BI)-Finanz-Architektur der Zukunft aus? Je Thema haben wir eine eigene Architekturskizze skizziert – auch wenn schlussendlich in jedem Finanzinstitut natürlich nur eine BI-Architektur gleichzeitig existieren kann. Unterschiede und Besonderheiten der einzelnen Anwendungsfälle lassen sich dadurch besser erkennen.

Im ersten Teil unserer Artikelreihe untersuchen wir, wie Sie Auswertungen in (fast) Echtzeit möglich machen und ungeahnte Datensätze für Ihr Unternehmen heben. In Teil II

werden wir uns mit Selfservice-BI beschäftigen und abschließend noch eine völlig neue Welt andeuten: eine aufsichtskonforme BI-Finanz-Architektur ohne (klassisches) Datawarehouse.

(Near-)Realtime – nie mehr Stress bei Ad-hoc-Prüfungen

Sie möchten einen Drill-Down in die Quartalsergebnisse schon während des morgendlichen Zähneputzens durchführen? Oder künftigen Ad-hoc-Anforderungen der Aufsicht, wie zum Beispiel der Fire-Drill-Übung der EZB¹, gelassener entgegengehen können? In-Memory und ein Kennzahlen-Repository machen es möglich.

1 Bei der Fire-Drill-Übung ist die Erwartungshaltung der Aufsicht an die Banken, gewisse Kennzahlen, beispielsweise die Liquidity Coverage Ratio und die damit zusammenhängenden Kennzahlen wie der CoCBC und der Maturity Ladder sehr zeitnah (t+1 oder zumindest t+2) berichten zu können (Quelle: Erfahrungsbericht der Autoren bei einer großen deutschen Landesbank).

Der Begriff „In-Memory“ ist schon seit einigen Jahren en vogue und darf in (beinahe) keiner Zukunftsprognose fehlen. Er beschreibt, dass Daten nicht mehr im (sehr langsamen) Plattenspeicher gehalten werden, sondern im ungleich schnelleren Arbeitsspeicher, englisch „memory“. Zum Vergleich: Eine Reise, die im Plattenspeicher etwa so lange dauern würde, wie ein Flug zum Pluto und zurück, entspricht im Arbeitsspeicher einer einstündigen Fahrt mit der U-Bahn. Und Arbeitsspeicher wird immer günstiger.

Oft taucht In-Memory in Begleitung der beiden Begriffe „Realtime“ und „Analytics“ auf. Das ist korrekt, gleichwohl möchten wir uns in diesem Kapitel auf „Realtime“ konzentrieren. Es geht also zunächst nicht darum, neue Datenschätze zu heben (das werden wir im Kapitel „Analytic Excellence“ diskutieren), sondern ausschließlich darum, die gewohnten Anfragen auf den Unternehmensdatenhaushalt in Echtzeit beantworten zu können.

Hierfür braucht es im Wesentlichen drei Dinge: erstens ein Datawarehouse (DWH) auf In-Memory-Basis, zweitens einen herkömmlichen Plattenspeicher und drittens ein Kennzahlen-Repository. Während das In-Memory-DWH, wie oben beschrieben, der enormen technischen Beschleunigung von Datenzugriffen dient, ist es die Aufgabe des Plattenspeichers, die DWH-Ergebnisse zu persistieren und damit gegen Ausfälle des Arbeitsspeichers zu sichern.

Das Kennzahlen-Repository beinhaltet die Definitionen aller relevanten Kennzahlen und ist der fachliche Motor zur Reporting-Beschleunigung. Bereits die Konzeption des Repositoriums hilft, ein institutswweit einheitliches Verständnis Ihrer Kennzahlen zu entwickeln (siehe Infobox). Die zentrale Ablage des Repositoriums ermöglicht es, Ad-hoc-Anfragen künftig schneller zu beantworten. Um aber tatsächlich alle Vorteile der In-Memory-Datenbank auszuschöpfen, müssen die Kennzahlen automatisiert berechnet werden können. Eine einfache Excel-Liste reicht daher nicht aus – das Kennzahlen-Repository muss die Formeln und DWH-Verknüpfungen (siehe Infobox) maschinenlesbar vor-

halten. Ob dabei auf bereits hochaggregierte DWH-Ergebnisse zugegriffen wird oder die Berechnung der Kennzahlen sogar auf Einzelgeschäftsebene im Kennzahlen-Repository erfolgt, ist gesondert zu entscheiden.

Natürlich sind auf einem In-Memory-DWH auch klassische Datenauswertungen über Datamarts möglich; ebenso können auch Quellsysteme mit In-Memory betrieben werden. Da wir aber davon ausgehen, dass ein zentraler Single-Point-of-Truth aus aufsichtsrechtlichen Gründen auch in 20 Jahren noch gefordert sein wird, beschränkt sich unsere Darstellung auf das DWH als essenzielle Komponente.

Analytic Excellence – entscheidend sind die Quelldaten

Zahllose Tools und Technologien zur Analyse von Daten sind in den vergangenen Jahren aus dem Boden geschossen. Doch die meisten dieser Tools setzen auf wohlstrukturierten Daten auf. Sie gehen also stillschweigend von einer bereits idealen Datenwelt des Anwenders aus. Die Realität hingegen zeigt: Vor allem der Weg der Datenbewirtschaftung muss grundlegend neu gedacht werden, um die Informationen aus den unendlichen Datenströmen nutzen zu können.

Kennzahlen-Repository und BCBS 239

Ein Kennzahlen-Repository ist weit mehr als nur eine Auflistung finanzmathematischer Formeln: Um wahren Nutzen zu stiften, umfasst es auch konkrete Zuordnungen der Kennzahl-Input-Parameter zu Feldern in Ihrem DWH. Fragt der Nutzer (oder die Aufsicht) beispielsweise die LCR-Kennzahl an, genügt ein Blick in das Kennzahlen-Repository, um zu wissen, welche Datenbankeinträge aufsummiert und durch welche anderen Werte geteilt werden müssen.

Dies erhöht einerseits den Komfort für den Nutzer bei häufig benutzten Anfragen, andererseits werden dadurch deterministische und folglich prüfungssichere Ergebnisse sichergestellt. Ein Kennzahlen-Repository ist daher nicht ohne Grund ein häufig eingesetztes Werkzeug in BCBS-239-Projekten.

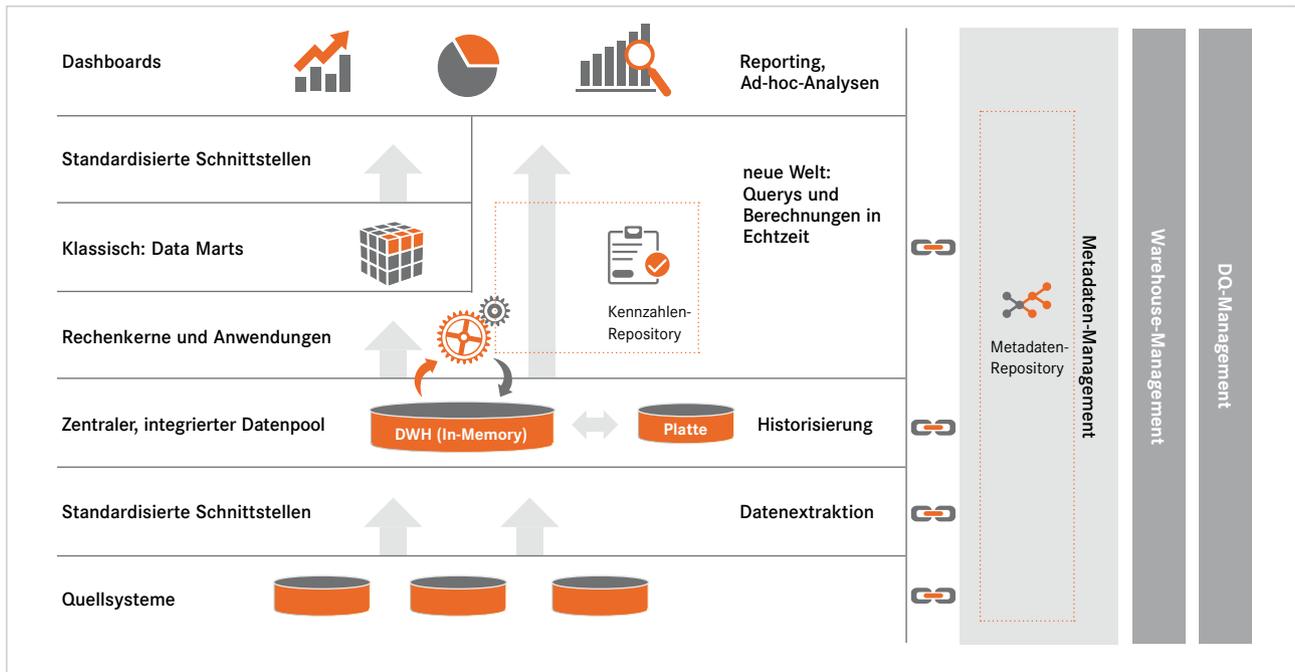


Abbildung 1: (Near-)Realtime: In-Memory und Kennzahlen-Repository ermöglichen Reports in (Fast-)Echtzeit.

Der grundlegende Unterschied von der heutigen zur zukünftigen Welt liegt in den Eigenschaften von „Streaming“-Daten. Streaming bedeutet, dass Daten (theoretisch) rund um die Uhr nach dem Push-Prinzip veröffentlicht werden. Reuters und Twitter sind zwei bekannte Beispiele für Unternehmen beziehungsweise Technologien, die Informationen streamen: Sobald es eine Neuigkeit gibt, wird sie an alle Empfänger verteilt. Die Krux dabei ist: Als Nutzer wissen Sie nie, welche dieser Informationen gerade für Sie Relevanz haben. Die erste Herausforderung besteht also für Ihr Unternehmen darin, die wesentlichen von den unwesentlichen Informationen zu unterscheiden. Dazu kommt, dass die Streaming-Dienste völlig unterschiedliche Datenformate vorweisen. Die Bandbreite reicht von csv-Dateien über Prosatexte aus Blogs bis hin zu Videoformaten. Folglich ist die zweite Herausforderung, diese Datenformate in die Struktur Ihres Datenhaushalts zu integrieren.

In einem ersten Schritt müssen die Daten von Twitter und Reuters in die Datenwelt Ihres Unternehmens gelangen. Da es sich um Daten handelt, die fortlaufend nach dem Push-Prinzip entstehen, sind Mechanismen notwendig, die fortwährend ihr „Ohr“ an diese Streams halten und die Daten aus den Streams in eine Datenbank schreiben, die unter Ihrer Datenhoheit steht. Diese

temporären Data Stores speichern die Daten nur so lange, bis sie weiterverarbeitet wurden. Für Ihre internen Bestandssysteme mit sogenannten In-situ-Daten, also Daten der „alten“ Welt mit klaren Strukturen und Gültigkeitszeitpunkten, können Sie entweder ebenso vorgehen oder einfach Ihre bestehenden Datenbewirtschaftungsprozesse weiternutzen.

Diese In-situ-Daten können dann entweder auf bekanntem Weg in Ihr DWH gelangen. Oder Sie lassen Teile daraus in Ihren Data Lake fließen. Für den Data Lake werden die Streaming-Daten sowohl technisch (zippen, ...) als auch inhaltlich komprimiert, beispielsweise nach Zeitscheiben aggregiert. Wichtig: Dafür muss bekannt sein, welche Daten aus den temporären Data Stores auf welche Weise komprimiert werden sollen. Sollen die Marktdaten auf Sekunden- oder gar auf Zehntel-Sekunden-Basis aggregiert werden? Können bei Blogbeiträgen die enthaltenen Grafiken gelöscht werden oder müssen sie für weitere Analysen erhalten bleiben? Bereits hier müssen daher weitreichende Entscheidungen getroffen werden und eine Anbindung an das Metadaten-Repository ist Pflicht.

Nun folgt die Königsdisziplin: die Strukturierung und Kombination der unprozessierten Rohdaten. Sogenannte Data-Scientists

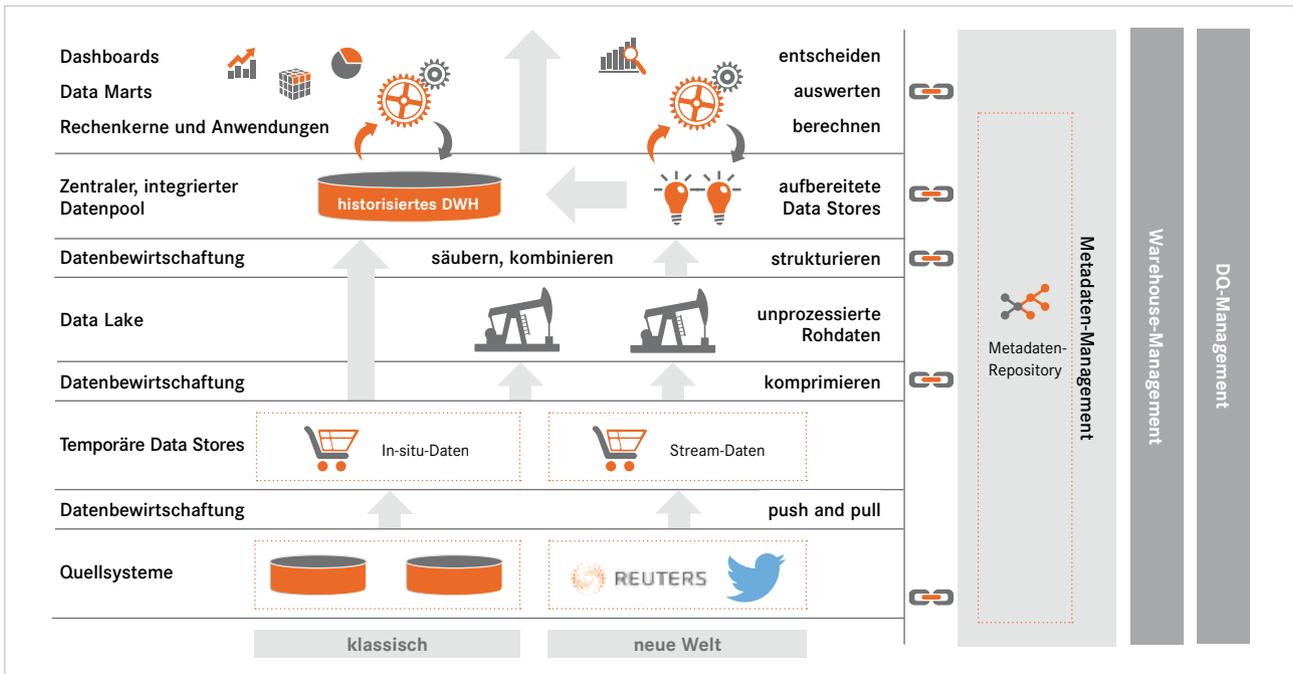


Abbildung 2: Analytic Excellence: Die Datenbewirtschaftung bis zum DWH verändert sich in der neuen Welt grundlegend.

definieren und selektieren Felder aus den Rohdaten. Sammeln Sie nur Datum und Name jenes Post-Verfassers, der auf Twitter Ihren Kundenservice kritisiert? Oder auch die weiteren Tweets des Verfassers und dessen Kontaktliste? Apropos Kontaktliste: Über Fremdschlüssel werden Tabellen miteinander verknüpft und Beziehungen hergestellt. Wie umfassend selektiert werden muss, hängt auch von der Leistungsfähigkeit Ihrer Datenbank ab – je leistungsfähiger, desto mehr (scheinbar nutzlose) Daten können Sie sammeln, die später einmal ungeahnten Nutzen bringen können ... Eine Säuberung der Daten, beispielsweise die Bereinigung um Redundanz oder unbrauchbare Meta-Informationen, kann möglicherweise auch schon im Schritt davor erfolgen.

Die derart aufbereiteten Daten verfügen nun über die gleiche Datenqualität wie die „klassischen“ DWH-Daten. Sie können entweder für Berechnungen in den Banksteuerungsanwendungen verwendet oder im DWH historisiert werden.

Im zweiten Teil unserer Artikelreihe lesen Sie:

- > Wie Sie mit Selfservice-BI (auch bekannt als Sandboxing) Ihren Fachbereichen neue Möglichkeiten der individuellen Datenauswertung eröffnen und welche entscheidende Rolle die IT dabei spielt.
- > Eine aufsichtsrechtskonforme BI-Finanz-Architektur ohne klassisches Data Warehouse – geht das denn überhaupt?

Ansprechpartner



Mathis Manz

IT Consultant

Sparkassen-Finanzgruppe

> +49 (0) 89 / 94 3011 - 0

> mathis.manz@msg-gillardon.de

Quellen/Referenzen: Pekka Pääkkönen, Daniel Pakkala (2015): Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. Big Data Research, Volume 2, Issue 4, Pages 166-186