Hybrides Data Warehouse am Beispiel von Data Vault

Shared Data in einer virtuellen Architektur

In der heutigen BI-Welt müssen Anforderungen der Fachbereiche und gesetzliche Vorgaben sowie die Wünsche der Nutzer von Smart Devices an personenbezogenen Daten in Einklang gebracht werden. Anhand des fiktiven Unternehmens FastChangeCo sollen im Folgenden die Herausforderungen an den Schutz der Daten sowie die Umsetzung aufgezeigt werden. FastChangeCo hat eine Möglichkeit entwickelt, nicht nur Smart Devices herzustellen, sondern auch die Smart Devices als Wearables in Form von Sensoren auf Kleidung auszudehnen. Das Geschäftsmodell von FastChangeCo hat nicht zum primären Ziel, die Technologie zu lizenzieren und gewinnbringend zu verkaufen. Im Gegenteil, das Ziel besteht darin, über niedrige Preise die Technologie möglichst breit im Markt zu positionieren um später mit einer großen Datenbasis Geld zu verdienen.

Hybride Data-Warehouse-Architektur

Immer häufiger berichten uns Familienmitglieder, Freunde und Bekannte stolz von eigenen Erfahrungen im Umgang mit Wearables wie Fitness-Trackern, Fitness-Apps, Smart Watches oder sogar vom vernetzten, intelligenten Heim. Bei jedem dieser Geräte entsteht eine große Menge an (sensiblen) Daten, genauer gesagt: durch die Aufzeichnung und Aufbereitung sowie die Auswertung personen- und umweltbezogener Daten. Dazu kommt, dass die kontinuierlich fortschreitende Entwicklung der Geräte einen ständigen Wandel der Schnittstellen zum Auslesen der Daten mit sich bringt [Coe15], [Fah15].

FastChangeCo steht daher nicht nur vor der Herausforderung, diese große Menge an Daten zu speichern oder die personenbezogenen Daten zu schützen, sondern auch davor, die agilen und flexiblen Anforderungen aus den Fachbereichen umzusetzen. Neben einer geeigneten Methode im Projektmanagement, zum Beispiel Scrum als agile Methode, und der physischen Modellierungsmethode, zum Beispiel Data Vault [Ler16], muss sich auch die IT der Rechenzentren den Anforderungen und Veränderungen, getrieben durch die Fachbereiche, stellen. Flexible und in kurzer Zeit umzusetzende Anforderungen an die Hardware-Infrastruktur können cloudbasierte Dienste meist sehr gut und dazu noch kostensparend lösen. Damit ist es der IT möglich, Rechenleistung temporär und innerhalb kurzer Zeit hinzuzubuchen.

Aufgrund der fachlichen, regulatorischen sowie flexiblen und agilen Anforderungen hat sich FastChangeCo für eine hybride Architektur entschieden, die Teile des Data Warehouse in der Cloud sowie lokal umsetzt. Um die gesammelten Daten auszuwerten, bedarf es einer Technologie, die es ermöglicht, den Fachbereichen eine effiziente, analytische Plattform zur Verfügung zu stellen, die eine uniforme Sicht auf die Daten ermöglicht.

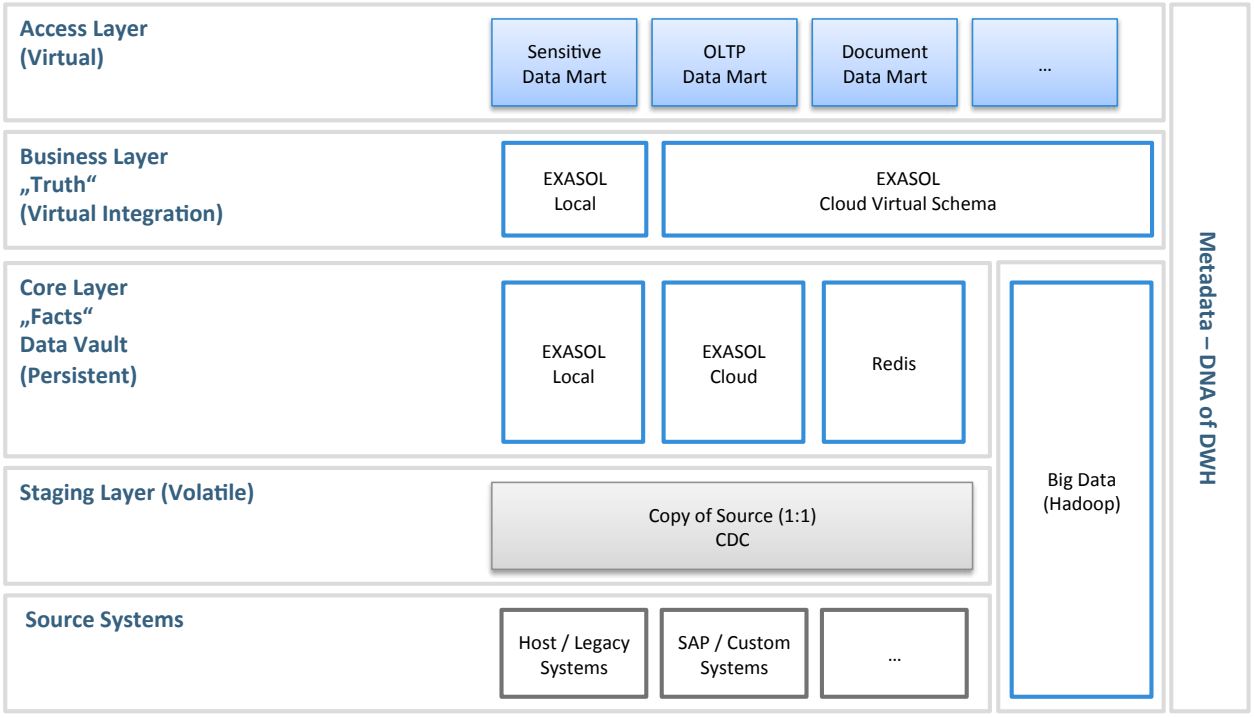

Abb. 1: Hybride Data-Warehouse-Architektur

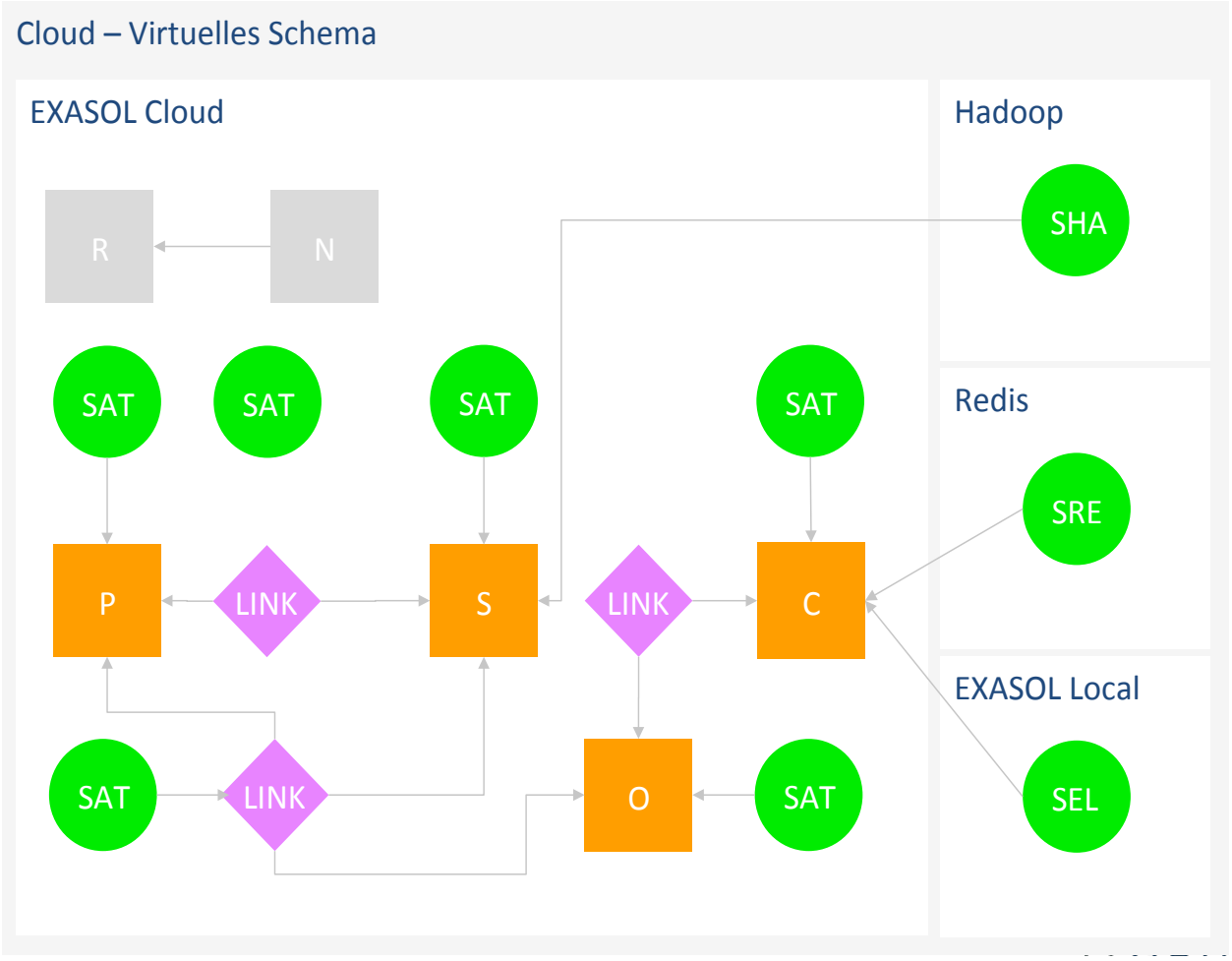

Abb. 2: Beispieldatenmodell FastChangeCo

Aufbauend auf der IT-Infrastruktur, die teilweise oder vollständig in die Cloud ausgelagert ist, sind hybride Ansätze, wie in Abbildung 1 dargestellt, zwischen Datenbanken in der Cloud, lokalen Datenbankinstallationen und weiteren Datenspeichern wie Hadoop oder NoSQL-Datenbanken notwendig, zum Beispiel MongoDB als Document Store oder Redis als Key Value Store.

Um die Daten zu einer logischen Gesamtsicht zu kombinieren, bedarf es daher einer geeigneten Methode der Datenmodellierung. Ein logisches Datenmodell bildet dabei die notwendigen Geschäftsobjekte sowie deren Attribute und Eigenschaften unabhängig von der eingesetzten Technologie eines Unternehmens ab. Wo, wie und mit welcher Technologie die anfallenden Daten des Data Warehouse persistiert werden, ist im logischen Datenmodell nicht relevant. Erst mit der Entwicklung der physischen Datenmodelle wird die eingesetzte Technik wichtig und somit die physische Modellierungsmethode. Beispielsweise bietet sich in dieser Konstellation die Modellierungsmethode Data Vault 2.0 an [Lin14], in der technologieübergreifend die Daten modelliert werden können, sowie eine Virtualisierungstechnik, die als zentrale Instanz alle beteiligten Technologien vereint. Der zentrale Aspekt der Virtualisierung ist, dass es für den Anwender unerheblich ist, wo die Daten tatsächlich liegen, der Anwender jedoch eine vollständig integrierte Datenlandschaft vorfindet.

Datenmodellierung

Die Firma FastChangeCo hat als Vorgabe, sensitive Daten unter der eigenen Kontrolle zu speichern und nicht in die Cloud auszulagern, das heißt, sie auf eigenen physischen Systemen im eigenen Rechenzentrum und von der eigenen IT zu verwalten. Trotzdem soll ein Zugriff auf die sensitiven Daten auch aus der Cloud heraus möglich sein, um die Daten im Bedarfsfall in ihrer Gesamtheit auswerten zu können.

Das in Abbildung 2 gezeigte Data-Warehouse-Datenmodell der Firma FastChangeCo basiert auf dem Datenmodell des TPC-H Benchmarks [TPC16]. FastChangeCo benutzt dieses einfache TPC-H-Datenmodell der 3. Normalform in seinen operativen Systemen und hat es nach seinen eigenen Bedürfnissen erweitert.

Das in Data Vault 2.0 erstellte Datenmodell des Core Data Warehouse ist in einer in der Cloud installierten Datenbank gespeichert. Dies stellt die zentrale Komponente des Data Warehouse dar. Zusätzlich wurde ein weiterer Satellit eingeführt, der physikalisch in einer Redis-Datenbank abgelegt ist und die eingescannten Verträge des Kunden beinhaltet („SRE“). Weiterhin wird ein Satellit des „Hub Supplier“ („S“) eingeführt. Dieser Satellit beinhaltet Maschinendaten in unstrukturierter Form und ist in einem HDFS abgelegt („SHA“). Um die personenbezogenen Self-Tracking-Daten

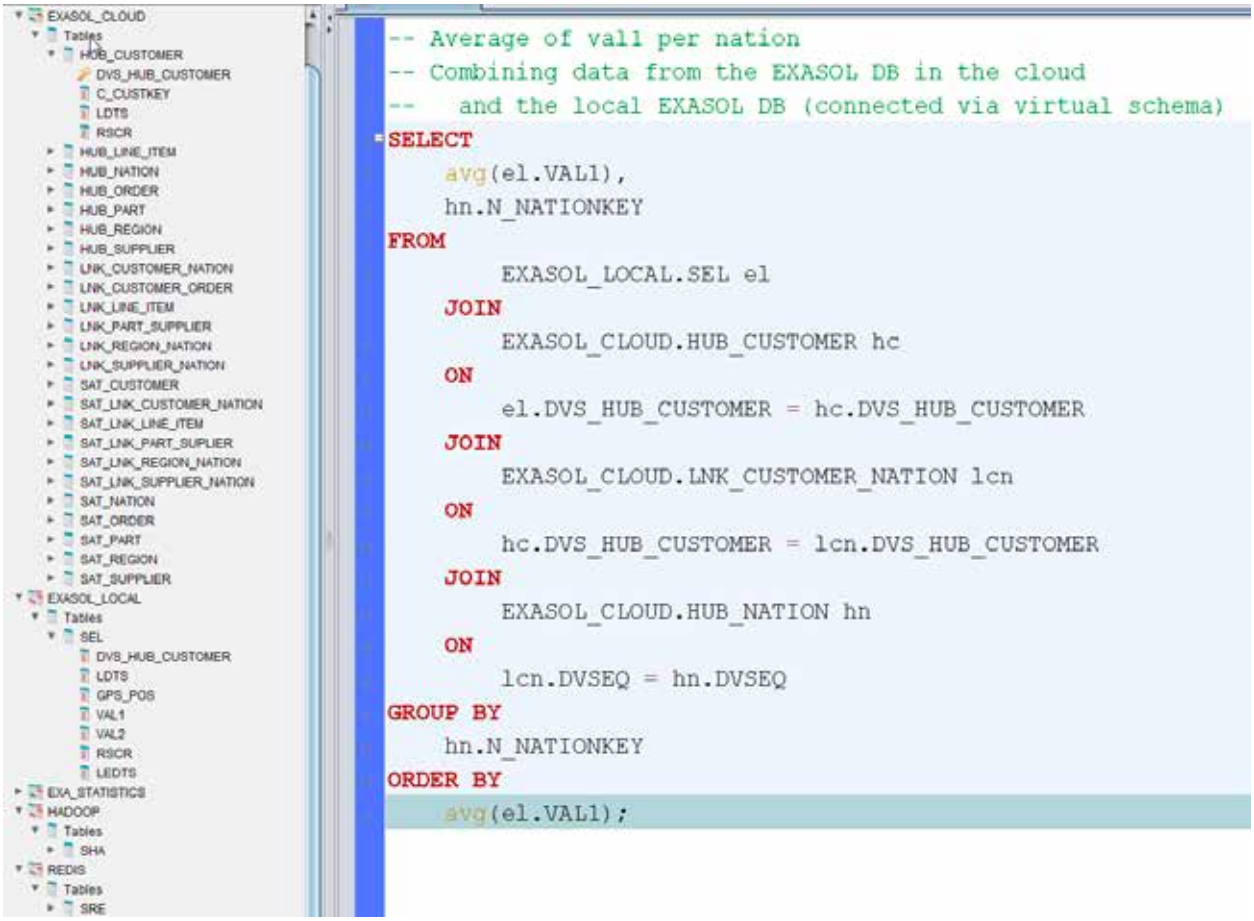

Abb. 3: Objekt-Browser mit virtuellen Schemata

unter höchsten Datenschutzmaßnahmen zu speichern, werden diese in einer lokalen Datenbank abgelegt. Im Datenmodell findet sich dazu der Satellit „SEL“.

Virtualisierung

Die Datenarchitektur aus Abbildung 1 kombiniert die Vorteile eines logischen Data Warehouse mit der Flexibilität der Data-Vault-Modellierung aus Abbildung 2. Als Virtualisierungstechnik und zur logischen Integration der unterschiedlichen Datenquellen wurden virtuelle Schemata in der Cloud-Datenbank angelegt. Diese ermöglichen einen transparenten Zugriff auf Daten, die in einer anderen relationalen oder NoSQL-Datenbank, aber auch in einem Hadoop-System abgelegt sind.

```
CREATE VIEW BusinessLayer.Customer AS
SELECT sel.a, sre.b, c
FROM ExaCloud.HubC hc
INNER JOIN ExaLocal.Sel sel
  ON hc.dvsk = sel.dvsk
INNER JOIN ExaLocal.Sre sre
ON hc.dvsk = sre.dvsk
```

Kasten 1: Beispielabfrage über lokale und virtuell eingebundene Tabellen

Alle nicht sensiblen und strukturierten Daten sind in der Cloud abgespeichert. Die verbleibenden Satelliten (unstrukturierte Maschinendaten der Lieferanten, eingescannte Verträge der Kunden, sensible Self-Tracking-Daten des Kunden) sind durch virtuelle Schemata angebunden.

In dem hier beschriebenen Beispiel bewegen sich Anwender, die keine sensitiven Daten sehen (dürfen), ausschließlich in der Cloud-Datenbank. Anwendern, die für sensitive Daten berechtigt sind, bewegen sich ebenfalls auf der Cloud-Datenbank, haben darüber hinaus aber auch Zugriff auf die lokale Datenbank und können somit transparent auf alle Daten zugreifen. Das Gleiche gilt für den Zugriff auf die Vertragsdaten in der Redis-Datenbank wie auch für die Maschinendaten in Hadoop. Abbildung 3 zeigt exemplarisch den Objekt-Browser mit allen virtuellen Schemata, wie es ein Anwender bei FastChangeCo mit vollständiger Berechtigung auf alle Datenbanken sehen würde.

Kasten 1 zeigt ein beispielhaftes SQL-Statement, wie FastChangeCo die Daten anhand der virtuellen Schemata unabhängig vom Speicherort selektieren kann und als einfache virtuelle Tabelle im Business Layer (siehe Abbildung 1) zur Verfügung stellt. Durch die Virtualisierung im Business Layer ist es für den Anwender unerheblich, wo die Daten des Data Warehouse persistiert sind.

Aufgrund der Abstraktion weg vom Data Vault hin zum virtuellen Business Layer ist der Zugriff auf die Daten des Data Warehouse für nachgelagerte Anwendungen transparent und „barrierefrei“. Ein weiterer Vorteil des virtuali-

sierten Business Layer ist die Entkopplung der dem Data Warehouse nachgelagerten Anwendungen von Änderungen im Data Warehouse, sowohl der physischen als auch der semantischen Strukturen.

Fazit

Die Kombination einer agilen Modellierungsmethode mit Best-of-Breed-Technologien ermöglicht es Unternehmen wie FastChangeCo, die Anforderungen von Fachabteilungen wie auch gesetzliche Regularien in einem flexiblen und agilen Umfeld umzusetzen. Durch den Einsatz von Konzepten zur logischen Kombination der Daten entsteht dem Endanwender kein zusätzlicher Aufwand, um die Daten zu vereinigen. Dies stellt neben dem Einsatz von Cloud-Technologien einen kostensparenden, aber hochperformanten Ansatz dar, der ohne größere Administrationsaufwände umsetzbar ist.

[Literatur]

[CoU15] Coester, N. / Ulla, P.: Autonomie hat oberste Priorität. In: BI-Spektrum, 5-2015, S. 08–11

[Fah15] Fahner, G.: Werden Körper mit Computern verschmelzen? In: BI-Spektrum, 5-2015, S. 22–24

[Ler16] Lerner, D.: Data Vault für agile Data-Warehouse-Architekturen. In: Agile Business Intelligence. dpunkt.verlag 2016

[Lin14] Linstedt, D.: #NoSQL platforms and #datavault curiosity, #bigdata and #datamodeling. 1.7.2014, http://danlinstedt.com/allposts/datavaultcat/nosql-platforms-and-datavault-curiosity-bigdata-datamodeling, abgerufen am 30.10.2016

[TPC16] TPC: siehe unter http://www.tpc.org/tpch/, abgerufen am 30.10.2016

BI-SPEKTRUM ist eine Fachpublikation des Verlags:
SIGS DATACOM GmbH | Lindlaustraße 2c | 53842 Troisdorf
Tel.: +49 (0) 22 41.2341-100 | Fax: +49 (0) 22 41.2341-199
E-mail: info@sigs-datacom.de
www.javaspektrum.de | www.objektspektrum.de
www.bi-spektrum.de

SIGS DATACOM
FACHINFORMATIONEN FÜR IT-PROFESSIONALS

Mathias Brink ist Solution Engineer bei EXASOL. Er berät und betreut Kunden und Interessenten beim Aufbau analytischer Plattformen im BI- und Big-Data-Bereich. **E-Mail: Mathias.Brink@EXASOL.com**

Dirk Lerner ist IT Senior Consultant bei ITGAIN und verantwortet dort als Teamleiter das Competence Center Data Architecture, Data Modeling & Data Vault. Seit rund 15 Jahren leitet er BI-Projekte und gilt als Experte für BI-Architekturen und Datenmodellierung. Herr Lerner ist ein Verfechter von flexiblen, schlanken und leicht erweiterbaren Data-Warehouse-Prinzipien und -Praktiken. Er ist ein Pionier für Data Vault in Deutschland und Autor des Blogs http://www.datavaultmodeling.de. **E-Mail: Dirk.Lerner@itgain.de**