tdwi Ausgabe 1 | 18. Jahrgang 2023 | D € 19,90 A € 23,90 SFR 35,25 ISSN 1862-5789

BI Spektrum

EINE PUBLIKATION DES TDWI E.V.

Datenplattformen: Worauf Sie bei der Einführung achten sollten und wo sich der Einsatz lohnt ab Seite 6

Datenintegration
Automatisierter und agiler Data Vault
Seite 26

Echtzeit-Architekturen
Zusammenspiel von MLOps und DevOps
Seite 32

Interview
Datenvirtualisierung bricht Silos auf
Seite 23

Alberto Pan, Denodo

zum Inhalt

Der Schatz im Silbersee: Agile Datenintegration eines Data Lake

Ein agiler und automatisierter Data Vault

Ein Beitrag von
Gunar Hofmann

In den letzten Jahren gab es bei vielen großen Unternehmen Initiativen, einen Data Lake als zentrale Plattform für den Datenzugriff im Bereich Analytics/Reporting/BI aufzubauen. Dabei führten Big-Data- und Data-Analytics-Projekte in einem ersten Schritt zur Etablierung einer Datenablageschicht basierend auf Hadoop oder S3. Darauf aufbauend wurden vielfältige meist spezialisierte Werkzeuge für einzelne Projekte, Anforderungen (unter anderem Streaming) und Fragestellungen implementiert.

Die Grenzen dieses Ansatzes sind schnell beschrieben: Um bei dem Bild des Data Lake zu bleiben, wird jedes Analytics/Reporting-Projekt in einem selbstgewählten Fischerboot ausgesetzt und darf mit einer Angel nach den Daten im See fischen, diese dann verarbeiten, veredeln und als fertige Auswertung auf den Markt bringen. Im besten Fall sprechen die Fischer noch miteinander und tauschen sich aus, wo es eine bestimmte Art von Fisch zu fangen gibt. Wir schauen uns deshalb einen alternativen Ansatz an, durch den projektübergreifend auf agile Art eine integrierte Datenschicht entstehen kann.

Wo stehen wir heute?

Der Ansatz des Data Lake erlaubt eine zentrale Beschaffung von operativen Daten für alle beteiligten Abnehmer und reduziert so die Zeit, bis Daten bereitstehen und ausgewertet werden können [SchoB20]. Er führt mehr Daten als bisher üblich an einem Platz zusammen und ermöglicht so kürzere Durchlaufzeiten für Datenanalysen [DBS18]. Das ist ein großer Fortschritt im Vergleich zu Data-Warehouse-(DWH-)Projekten, bei denen allein die Abstimmung der zu liefernden Daten und die Bereitstellung durch die operativen Systeme schon

Bild: Shutterstock

zum Inhalt

Monate benötigen konnte. Die Analysen auf einem Data Lake basieren auf denselben operativen Rohdaten, was ein weiterer Fortschritt im Vergleich zu lokalen Auswertungen von operativen Daten darstellt (vgl. Abbildung 1).

Von einem Data Lake in eine Silo-Landschaft

Die unterschiedlichen Projekte auf einem Data Lake verwenden häufig ganz verschiedene Softwareplattformen und erzeugen keine gemeinsamen Datenstrukturen, sondern implementieren diese für jeden Anwendungsfall [Hay20]. Für eine BI- oder Reporting-Lösung oder auch die Abbildung eines Geschäftsprozesses ist dieser Ansatz als Data Silo [Foo21] bekannt und hat sich aufgrund fehlender Wiederverwendbarkeit, uneinheitlicher Datenqualität, der Gesamtkosten und des Wartungsaufwands bei einer langen Lebenszeit nicht durchgesetzt. Ähnliche Geschäftslogik wird so mehrmals definiert und implementiert. Geschäftsobjekte des Unternehmens können je nach Projekt ganz unterschiedliche Ergebnisse liefern. Falls die geplante Lebenszeit dieser Lösungen nur wenige Jahre beträgt, wie im ersten Schritt häufig bei Analytics-Projekten, kann der Ansatz sinnvoll sein, um kostengünstig und schnell den Erfolg eines Ansatzes bewerten und implementieren zu können. Data Silos an sich sind keine neue Herausforderung und wurden seit den 80er-Jahren durch zentrale DWH-Systeme für die Datenintegration, Qualität, Security und Governance gelöst [BaK21], und diese Herausforderung stellt sich nun erneut.

Das Data Lakehouse als DWH auf dem Data Lake

Ein weiteres aktuelles Angebot am Markt beschreibt einen Satz von Werkzeugen, der als Data Lakehouse von fast allen großen Cloud-Anbietern aktuell gepusht wird [Hel21]. Damit kann zumindest die technische Plattform harmonisiert werden. Die einzelnen Softwarelösungen ähneln frappierend denen aus vergangenen Data-Warehouse-Projekten und lassen zumindest den Schluss zu, dass Themen wie Datenqualität, Governance, Scheduling weiterhin essenziell wichtig sind.

Sehr ähnlich zu den Werkzeugen für ein Data Warehouse, wie zum Beispiel einer ETL/ELT-Software, fängt eine Implementierung praktisch bei null an, was viele Freiheiten lässt, aber auch sehr zeitaufwendig sein kann. Ein neuer Modellierungsansatz für ein übergreifendes Datenmodell ist nicht zu erkennen und fertige Softwaremodule für bestimmte Aufgaben sind beim Data Lakehouse aktuell nur in Anfängen implementiert (siehe Abbildung 2).

Der Data Vault als Ansatz zur Integration von Daten

Der Data Vault als alternativer Modellierungsansatz, der inzwischen gut mit Softwarelösungen abgebildet wurde, bietet hier einen Standard zur Mo-

Abb. 1: Der Data Lake und seine Nutzung am Beispiel AWS (Quelle: [ALF22])

GUNAR HOFMANN ist Gründer und Geschäftsführer der syntegris information solutions GmbH, arbeitet seit seinem Abschluss als Diplom-Wirtschaftsinformatiker 1997 als Consultant in Projekten mit relationalen Datenbanken, seit 2000 als Berater, Designer und Architekt in Business-Intelligence- und Data-Warehouse-Projekten und ab 2020 in Data-Vault-Projekten in der Implementierung mit Hilfe verschiedenster Data-Vault-Softwarelösungen. Er ist seit 2002 als Praxispartner für das duale Studium Wirtschaftsinformatik aktiv und vergibt und betreut Bachelor- und Masterarbeiten im Bereich DWH/BI.
E-Mail: datavault@syntegris.de

zum Inhalt

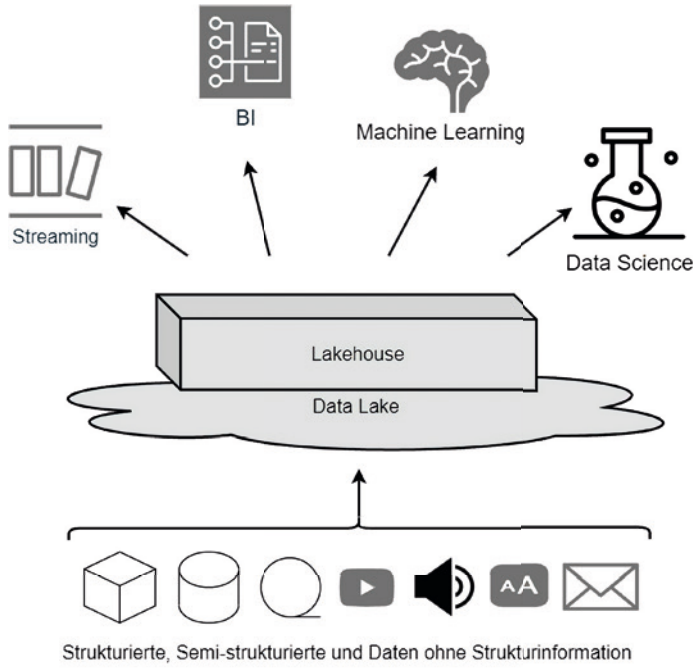

Abb. 2: Übersicht über ein Data Lakehouse

dellierung der benötigten Integrationsschichten. Es gibt eine klare Strategie, wie die Modellierung iterativ erfolgen kann [LiO15].

Als Open-Source-Software-Implementierung sind hier das Virtual Data Warehouse von Roelant Vos [Roe22] und dbtvault [DBT22] zu nennen. Im Folgenden wird die Data-Vault-Methode anhand der Software Datavault Builder vorgestellt, auch wenn dies in stark gekürzter Form erfolgen muss [DVB22].

Ein Data Vault bietet gegenüber einem Data Lakehouse erhebliche Vorteile (vgl. Abbildung 3):

- Er kann modelliert und die meisten Transformationen können automatisch generiert werden.
- Er bildet den „single point of truth“, wie bei einem DWH
- Prüfung und Steigerung der Datenqualität
- Integration verschiedener Datenquellen über gemeinsame Business Keys
- Bottom-up-Entwicklung des Datenmodells
- „Start small“ mit agilen Projekten
- Unterstützung von DevOps
- Wiederverwendung von Dimensionen und Kennzahlen
- Ein gemeinsamer Business Vault ist möglich, aber nicht zwingend vorgeschrieben
- Die Anpassung oder der Ersatz von Datenquellen ist deutlich einfacher als in einem DWH
- Reporting Layer über Business Rules und Dimensional Model

Die ersten Phasen beim Aufbau eines Data Vault

Im Gegensatz zum Data Lake, bei dem die Daten in beliebigen Formaten, Strukturen oder auch als Text abgelegt werden, wird im Data Vault den Daten als Erstes eine definierte Struktur zugewiesen und so ein Raw Vault befüllt. Beim Start der Nutzung der Plattform kann jedes Projekt unabhängig voneinander Daten anbinden, eine eigene Subject Area im Raw Vault anlegen und Auswertungen darauf erstellen. Sobald es Überschneidungen zwischen Projekten gibt, müssen die gemeinsamen **Hub-Objekte** und deren **Business Keys** harmonisiert werden, eine Angleichung der Datenmodelle ist nicht nötig, aber sinnvoll.

Auslesen des Data Lake erfolgt über die Staging-Schicht

Als Erstes werden die Daten in eine Stage-Schicht geladen, dadurch erhalten alle Attribute einen definierten Datentyp. Die Datenqualität der Attributwerte kann verbessert und über 1:1-Mappings eine Harmonisierung der Attributwerte erzielt werden.

Der Raw Vault als Langzeitspeicher und Integrationsschicht

Im Raw Vault, der **ersten zentralen** Schicht, werden alle Datenstrukturen in Hubs, Links und Satelliten zerlegt. Diese Schicht ermöglicht die Integration der Datenquellen und sie kann flexibel auf Änderungen der Datenquellen reagieren. Die Hubs in einem Raw Vault sind die **einzigen unabhängigen Datenstrukturen** und werden durch einen **Business Key** gebildet. Die Beziehungen zwischen den Hubs werden durch Links abgebildet, die eine flexible Verbindung zwischen den Hubs darstellen (n:m-Kardinalität). Sonstige Attribute aus einer Datenlieferung werden in Satelliten gespeichert und den Hubs zugeordnet.

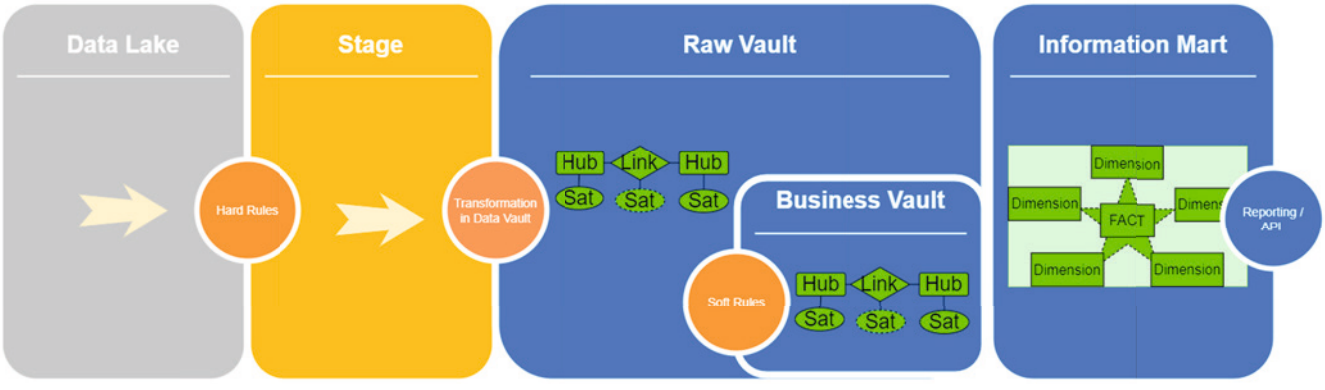

Abb. 3: Schichtenmodell des Data Vault

zum Inhalt

Hubs als unabhängige zentrale Anker

Das Konzept des Hubs ist zentral für den Raw Vault und die Definition des Business Key eines Hubs unterscheidet sich stark von der Definition eines Primärschlüssels in einem relationalen Modell. Ein Business Key **muss** ein im Unternehmen **verwendeter** Fachschlüssel sein, der in den Anwendungen **sichtbar** ist und **zentral** für den zu bildenden Hub im Unternehmen steht. Analog zu einem Primary Key muss der Business Key in einem Hub **einzigartig** und **immer gefüllt** sein. Technische Schlüssel und Surrogate Keys sind nur in seltenen Fällen als Business Keys zu verwenden. Man kann auch sagen, nachdem die natürlichen Schlüssel sich in der relationalen Modellierung nicht durchsetzen konnten, sind sie nun wieder im Data Vault zurück. Beispiele für einen Business Key sind: IATA-Airline-Codes von Fluglinien, für Banken im SEPA-Raum kann es die BIC sein, oder ein Telefonanschluss wird über seine Landesvorwahl, Ortsvorwahl und Rufnummer identifiziert. Die IBAN eines Bankkontos, die in großen Teilen Europas und Afrikas verwendet wird, ist ebenfalls ein Business Key. Wenn Sie sich im Internet bei einer Plattform anmelden, kann die URL zusammen mit dem Accountnamen den Business Key für einen Hub bilden. Bei Flügen ist zum Beispiel der Airlinecode, die Flugnummer des Operators, also der ausführenden Fluglinie des Fluges, und eine laufende Nummer der Business Key.

Links bilden Beziehungen ab

Die Links als zweites Element des Raw Vault verknüpfen Hubs miteinander. Sie dienen der Modellierung einer **konkreten** Beziehung. Beispiele für einen Link zwischen Hubs sind: Eine Fluglinie führt Flüge durch oder ein Kunde besitzt ein Bankkonto. Es sind auch mehrere Links zwischen Hubs modellierbar.

Satelliten als Attributspeicher

Als drittes Element im Raw Vault werden Satelliten gebildet. Wenn Sie sich fragen, wo denn nun die vielen **Attribute der Datenquelle** gespeichert werden, welche man später auswerten möchte und die **kein** Business Key sind, dann werden Sie diese in einer Reihe an Satelliten im Raw Vault finden. Satelliten werden nach technischen Kriterien erzeugt, zum Beispiel je nach Datenquelle und Häufigkeit der Änderung der Daten im Satelliten. Zu einem Hub kann jederzeit ein neuer Satellit gebildet werden. Mit den Satelliten lassen sich neue Datenquellen, Änderungen in Datenquellen sowie die Historisierung von Daten einfacher abbilden und behandeln.

Komplexe Geschäftslogik wird im Business Vault implementiert

Aufbauend auf dem Raw Vault werden im Business Vault alle Geschäftsobjekte abgebildet, die Daten aus mehreren Quellen verwenden oder neue Daten aus den Datenquellen bilden. Die Struktur der Geschäftsobjekte wird weiterhin mit Hubs, Links und Satelliten abgebildet. Hier ist auch der richtige Ort für komplexe Logiken, die programmiert beziehungsweise über KI oder Analytics erzeugt werden. Für ein zentrales Reporting-Backend ist der Business Vault zusammen mit dem Raw Vault die richtige Datenbasis, auf weitere Vaults wie den Error Vault oder den Metric Vault können wir hier nur hinweisen. Darauf aufbauend werden Dimensionen, Kennzahlen und Data Marts gebildet (vgl. Abbildung 4).

Starten Sie klein und agil

Der Modellierungsansatz ermöglicht eine agile und iterative Entwicklung, weil Sie direkt mit den ersten zu verarbeitenden Daten starten können. Das Datenmodell baut sich anhand der zu verarbeitenden Daten auf. Es werden am Anfang einzelne Inseln entstehen, die erst später über weitere Hubs verbunden werden. Es ist keine Designphase für ein übergreifenden Datenmodell nötig. Jedes einzelne Projekt kann so schnell wie möglich seine Daten über den Data Lake anbinden, in den Raw Vault integrieren und dem Reporting bereitstellen. Während die Hubs von mehreren Datenquellen befüllt werden können, wird jede Datenquelle separate Satelliten erzeugen, was das parallele Laden der Daten und unabhängige Anpassungen ermöglicht. Unter der Voraussetzung, dass die Daten auch im Data Lake vollständig vorhanden sind, können die Durchlaufzeiten im Vergleich zu einem DWH drastisch reduziert werden.

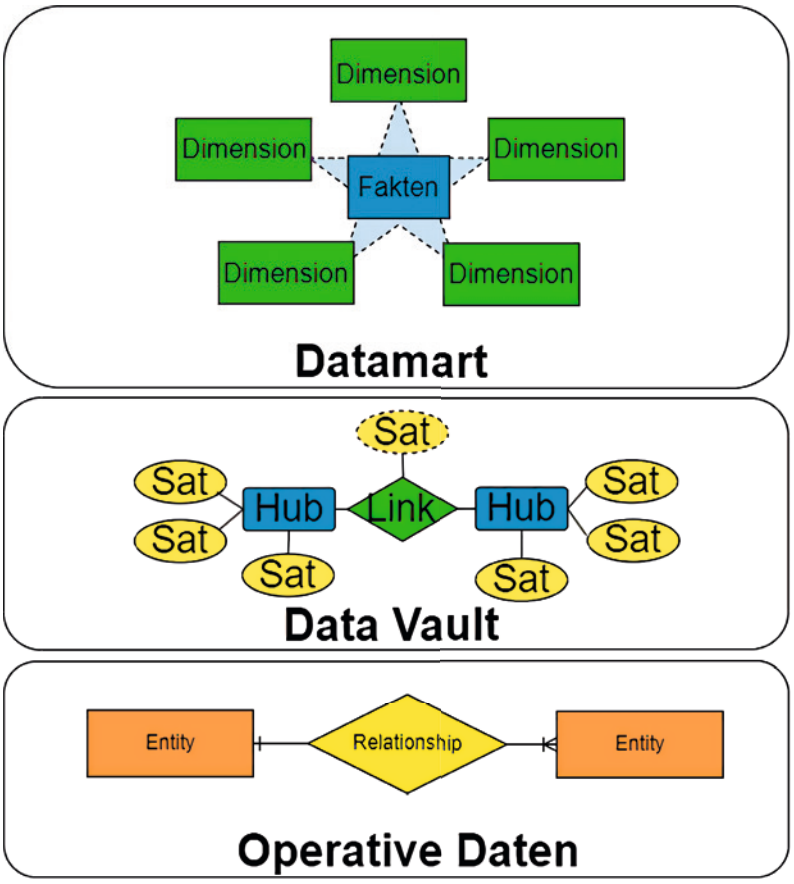

Abb. 4: Der Data-Vault-Ansatz: Teile und herrsche

zum Inhalt

Für eine Erweiterung oder Anpassung des Data Vault sind so im DevOps-Umfeld wenige Tage und im Scrum-Umfeld ein Sprint realistisch – das umfasst die Implementierung vom Data Lake bis zum Reporting. Hierbei ist aber die Verwendung einer Softwarelösung wie dem Datavault Builder vorauszusetzen.

Ideen für die Zukunft

Auch wenn ein Großteil der im DWH individuell erstellten Funktionen bereits im Datavault Builder vorhanden ist, bleiben noch weitere Themen für eine Unterstützung oder Automation offen. Für die Zukunft sehe ich folgende Anforderungen, die in einem Data Lake oder DWH häufig nicht abgebildet sind:

- **Vernetzung** der Akteure auf dem Data Vault, Teilen von Know-how und Datenstrukturen
- **Monitoring** der Datenaktualität, der Datenqualität und Anzeige von Veränderungen
- **Prüfung der Konsistenz** der Daten über Datenquellen hinweg
- **Prüfung der Integrität** des Raw Vault und des Business Vault
- **Automatisierung der Strukturerkennung**, wenn der Data Lake keine Metadaten zur Struktur liefern kann
- **Profiling** der Daten von neuen Datenquellen und Vergleich mit den Datenattributen im Raw Vault, um gleiche oder ähnliche Daten zu erkennen
- **Lebenszyklus** von Datenmodellen und den zugehörigen Logiken

Unsere Erfahrungen und das Fazit

Der Data Vault bietet Lösungen für eine Reihe an technischen Herausforderungen bei der Integration von Daten. Er ersetzt nicht die fachlichen Vorgaben bezüglich der zu modellierenden Business-Objekte und deren Harmonisierung innerhalb eines Unternehmens.

Der Data Vault als Alternative zu klassischen DWH-Implementierungen bietet konkrete Werkzeuge für die typischen Anforderungen an ein Reporting Backend an, wie die Integration von Datenquellen und die Bildung einer übergreifenden Business-Schicht. Dimensionen und Data Marts können in der Software modelliert und automatisiert implementiert werden. Die Software Datavault Builder beispielsweise erzeugt aus dem Raw-Vault- und dem Business-Vault-Modell direkt ein Datenbankmodell und lauffähige Verarbeitungsprozesse. Die Verwendung einer Source-Code-Verwaltung wie GIT sowie die agile Weiterentwicklung und Automatisierung des Deployments über CI/CD werden von der Software unterstützt. Für einen stabilen Raw Vault bleibt die übergreifende Definition der Business Keys der Hubs das zentrale Element: Man muss sich bei jedem einzelnen neuen Hub sicher sein, dass er aus Sicht des Gesamtunternehmens vollständig definiert ist und wiederverwendet werden kann. Der nachträgliche Austausch des Business Key eines Hubs ist zeitaufwendig, weil damit alle abhängigen Links und Satelliten angepasst werden müssen.

Der Raw Vault (Abbildung 5) sollte **nicht** von Analysten oder Endanwendern direkt verwendet werden. SQL-Abfragen auf der Datenstruktur sind

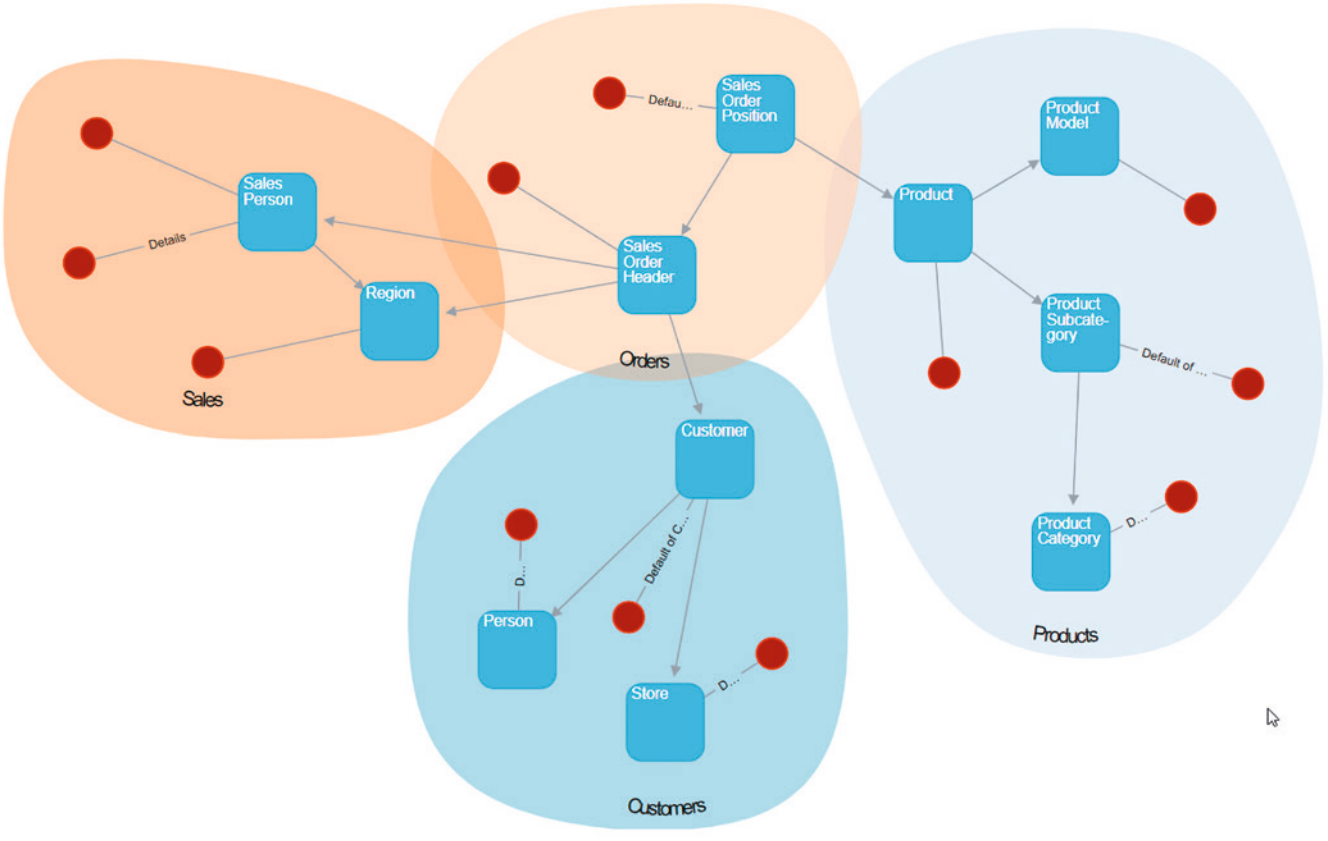

Abb. 5: Ausschnitt aus dem Raw Vault [DVB22]

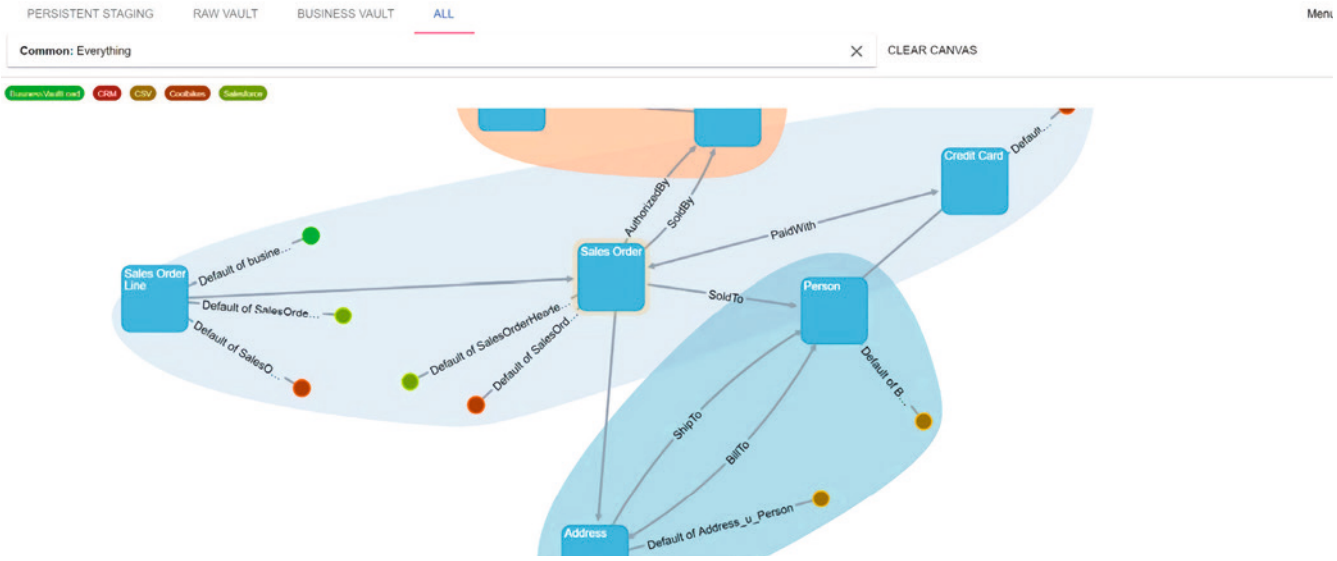

Abb. 6: Iterative Integration von Datenquellen im Raw Vault [DVB22]

mühsam zu erstellen und jede Änderung des Raw Vault kann Anpassungen in den Abfragen verursachen. Besser eignen sich hier der Business Vault oder die Data Marts im Reporting Layer.

Die Designprinzipien und die Strategie des Business Vault sind aktuell noch zu wenig standardisiert und in der Literatur wird dieser nicht detailliert beschrieben. Ob der Business Vault in der Struktur eines Data Vault optimal für Auswertungen und Analysen ist, muss sich noch zeigen.

Bei der Verarbeitung von unterschiedlichen Datenstrukturen und Formaten hat sich der Datavault Builder als sehr robust erwiesen. Ob relationale Strukturen, Prozessdaten, Graphen und Streams – die größte Hürde ist hier, einen Konverter für das Anlanden der Daten in der Stage zu finden oder zu entwickeln.

Grenzen des Ansatzes und der Software zeigen sich, wenn aggregierte oder denormalisierte Daten keinen Bezug zu bereits vorhandenen Hubs besitzen, weil der Business Key nicht Teil der Lieferung ist. Für einzelne Hubs können auch mehrere konkurrierende Kandidaten für den Business Key existieren. Bei der Konvertierung relationaler Strukturen darf man sich nicht dazu verleiten lassen, jeden Primärschlüssel in einen Hub zu verwandeln.

Aufgrund der Kürze des Artikels konnte ich nicht auf Details der Modellierungsmethode Data Vault und der Software Datavault Builder eingehen, sondern nur einen ersten Teil des Modellierungsprozesses und zentrale Vorteile skizzieren. Der Data Vault ist aktuell der einzige mir bekannte Modellierungsansatz für die Datenintegration, der sich in großen Teilen automatisieren lässt, agil entwickelt werden kann und Anpassungen an den Datenquellen durch Erweiterungen des vorhandenen Datenmodells abbildet (siehe Abbildung 6).

Literatur

[ALF22] AWS Lake Formation, https://aws.amazon.com/de/lake-formation/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc, abgerufen am 10.2.2023

[BaK21] Baars, H. / Kemper, H.G.: Business Intelligence & Analytics – Grundlagen und praktische Anwendungen: Ansätze der IT-basierten Entscheidungsunterstützung. Springer Vieweg 2021

[DBS18] DBSystel, zero.one.data: Bahn frei für Big Data. Juni 2018, https://www.dbsystel.de/dbsystel/Digital-Stories/Bahn-frei-fuer-Big-Data-6165578, abgerufen am 29.7.2022

[DBT22] dbtvault by Datavault, https://github.com/Datavault-UK/dbtvault, abgerufen am 10.2.2023

[DVB22] Reporting: Datenintegration und Vorbereitung – Flexibel. Skalierbar. Zugänglich. https://datavault-builder.com/de/use-case-datawarehouse-for-reporting/, abgerufen am 29.7.2022

[Foo21] Foote, K.: A Brief History of Data Silos. 7.10.2021, https://www.dataversity.net/a-brief-history-of-data-silos/#, abgerufen am 10.2.2023

[Hay20] Hayler, A.: Common data lake challenges and how to overcome them. 17.4.2020, https://www.techtarget.com/searchdatamanagement/feature/Common-data-lake-challenges-and-how-to-overcome-them, abgerufen am 10.2.2023

[Hel21] Heller, M.: Was ist ein Data Lake? 6.5.2022, https://www.computerwoche.de/a/was-ist-ein-data-lake,3553264, abgerufen am 10.2.2023

[Lin015] Linstedt, D. / Olschinke, M.: Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann Publ. Inc. 2015

[Roe22] RoelantVos/virtual-data-warehouse, 25.1.2022, https://github.com/RoelantVos/Virtual-Data-Warehouse/releases/, abgerufen am 10.2.2023

[ScB20] Scholz, O. / Bauer, S.: Cloud Transformation bei der Commerzbank. Oktober 2020, https://www2.deloitte.com/de/de/pages/financial-services/articles/google-cloud-transformation-commerzbank.html, abgerufen am 10.2.2023

zum Inhalt