Aufbau eines agilen, automatisierten Data Warehouse für SAP

SAP-Greenfield-DWH mit Data Vault

C&A, eines der führenden Modeunternehmen in Europa, führt aktuell eine SAP-Retail-Landschaft ein. Ziel ist es, die zahlreichen Geschäftsprozesse zu digitalisieren und zu optimieren. Der Bedarf für Analytics im Unternehmen ist hoch, sodass der Aufbau eines Greenfield Data Warehouse als Basis für Analytics, BI und Reporting von Anfang an Bestandteil des SAP-Programms wurde. Anders als in traditionellen SAP-Umgebungen hat sich C&A für die Einführung einer Best-of-Breed-DWH-Architektur auf Grundlage von Data Vault, Data Warehouse Automation (DWA) und einer Data-Lakehouse-Architektur entschieden, die auf Basis von AWS mit der Data Platform Snowflake und der DWA-Lösung Datavault Builder realisiert wurde.

Ein Beitrag von Lutz Bauer und Till Sander

Übergreifendes Ziel des Greenfield-DWH ist die Bereitstellung einer Data Platform für sämtliche Dateninhalte von C&A, das heißt sowohl aus der SAP-Retail-Landschaft als auch aus sämtlichen Non-SAP-Systemen. Dadurch sollen flexible und performante Möglichkeiten für Analytics und Reporting geschaffen werden. Eine einfache Datenintegration mit Non-SAP-Daten ist ein wichtiges, langfristiges Ziel, ebenso wie die direkte Verknüpfbarkeit von SAP-Daten mit Rohdaten (zum Beispiel Kassendaten, Sensordaten, Customer Counting, RFID) für Analytics.

Nach einem intensiven Auswahlprozess fiel die Wahl für die Data Platform auf eine flexible Best-of-Breed-Lösung basierend auf dem cloudbasierten Anbieter Snowflake.

Ein weiteres wichtiges Ziel ist es, eine zu SAP konsistente Business-Sicht des DWH-Systems auf die SAP-Landschaft zu bilden. Hierzu soll im DWH-System keine (oder nur geringfügig) Geschäftslogik abgebildet werden. Hinsichtlich der Geschäftslogik wird SAP als führend gesehen. Aus diesem Grunde traf C&A bereits früh die Entscheidung, die SAP BW Data Sources als Extraktionsquelle zu nutzen, da die Datenextraktion aus dem SAP-ERP bereits eine auswertungsnahe Datensicht bietet und Geschäftslogik nur in Ausnahmefällen im DWH aufzubauen ist.

Bild: Shutterstock

LUTZ BAUER ist seit 2019 als Domain Architect für Data und Analytics bei der IT-Tochter der C&A tätig. Seine fachlichen Schwerpunkte sind Data Integration, DWH-Plattformen und die Modernisierung von DWH-Architekturen. Er blickt auf über 25 Jahre Erfahrung im Bereich Data Warehousing und BI zurück. Er hat bereits als Autor in BI-Spektrum veröffentlicht und war mehrfach Sprecher auf der TDWI Konferenz und anderen Veranstaltungen.
E-Mail: lutz.bauer@canda.com

TILL SANDER weist eine fast 20-jährige Erfahrung als Manager und Lösungsarchitekt bei Technologie- und Consulting-Unternehmen auf. Vor areto war er als Solution Manager bei verschiedenen Beratungsunternehmen tätig. Als Chief Technical Officer (CTO) bringt er unter anderem seine langjährige Expertise in der Konzeption und dem Aufbau von Business-Intelligence-Lösungen in die Geschäftsführung ein. Auf dieser Basis treibt er den Auf- und Ausbau der areto consulting gmbh, die Evaluierung neuer Softwareprodukte sowie die Weiterentwicklung bestehender Service-Angebote und Beratungsleistungen weiter voran. Seine Ausdauer und seinen Willen, immer das Optimum zu erreichen, beweist er nicht nur in Kundenprojekten, sondern auch als passionierter Rennradler.
E-Mail: till.sander@areto.de

Weiterhin sollen über den agilen Ansatz – in kurzen, iterativen Umsetzungszyklen – produktive Abschnitte zur Verfügung gestellt werden. Für das Modeunternehmen C&A fiel die Wahl dabei auf einen Ansatz mit Data Warehouse Automation basierend auf Data Vault und dem Datavault Builder der 2150 Datavault Builder AG. Zusätzlich wurde eine Data-Lakehouse-Architektur in Verbindung mit Automatisierung eingeführt, um Aufwände bei späteren Refactoring-Aktivitäten einzusparen. Für die BI-Layer wird das bereits im Unternehmen etablierte ROLAP-Tool MicroStrategy eingesetzt. Für Analytics kommen Dataiku sowie verschiedene weitere Data-Science-Tools zum Einsatz.

Gewählter Ansatz: (Weitgehende) Automatisierung

Als Grundprinzip des C&A Greenfield Enterprise Data Warehouse steht die weitgehende Automatisierung bei der Entwicklung der einzelnen Schichten im Vordergrund (Abbildung 1). Wir fokussieren uns im vorliegenden Artikel auf die Quelle SAP.

Data Lakehouse: Das SAP-ERP schaltet die für das DWH relevanten Schnittstellen „Standard BW Extraktoren“ frei. Die ETL-Extraktion wurde so aufgebaut, dass die verfügbaren Schnittstellen automatisch erkannt werden und in der täglichen Batch-Extraktion automatisiert in das Snowflake-basierte Data Lakehouse im semistrukturierten Rohdatenformat JSON geladen wird. Sobald vom SAP-System neue oder geänderte Schnittstellen zur Verfügung gestellt werden (zum Beispiel nach einem SAP-Customizing), wird dies vom ETL-System „Ab Initio“ automatisch erkannt und entsprechend als Erweiterung in das Data Lakehouse geladen. JSON bietet hier eine Möglichkeit, um resilient gegenüber Schnittstellenänderungen zu sein.

Staging Preparation Layer: Aufgabe dieses Layers ist es, die in Snowflake gespeicherten JSON-Dateien mit Hilfe von SQL-Views zur Verfügung zu stellen. Weiterhin wird das von SAP angelieferte Change-Data-Capture-Format (Delta-Extraktion) so aufbereitet, dass die folgenden ETL-Prozesse damit effizient arbeiten können. Konkret heißt das: Begrenzung der Änderungshistorie auf den jeweils gültigen Stand zum DWH-Ladezeitpunkt. Durch einen Automatisierungsmechanismus wird für jeden neu angebundenen SAP BW Extraktor eine SQL-View erstellt. Der eingesetzte Code-Generator verwendet die Dictionary-Metadaten des Systems SAP, um die entsprechenden Zugriffs-Views automatisch zu erzeugen.

Data Vault Layer: Der Aufbau des integrierten Data Vault 2.0 [LiO15] erfolgt über das Data-Warehouse-Automation-(DWA-)Tool Datavault Builder. Zur Definition des logischen Datenmodells sind einige wenige Informationen anzugeben (zum Beispiel fachlicher Schlüssel, Link-Verbindungen). Das abgeleitete physische Datenmodell sowie die ETL-Logik werden vom DWA-Tool automatisiert erzeugt.

Die SAP-BW-Extraktoren orientieren sich an einzelnen Geschäftsobjekten (zum Beispiel Geschäftspartner, Transportauftrag, Sales Order). Als Modellierungs-Guideline wurde festgelegt, den Data Vault anhand der Darstellung der Geschäftsobjekte gemäß Sichtweise der SAP-BW-Extraktoren aufzubauen. Diese Entscheidung hat sich als sinnvoll erwiesen.

Der Business-Layer des Data Vault wird manuell definiert. Das physische Datenmodell und die Logik für Berechnungen, wie zum Beispiel Point-In-Time-Tabellen, werden vom Tool erzeugt.

Presentation Layer: Der Presentation Layer wird nach dem Prinzip der Dimensionalen Modellierung durchgeführt [KiR13]. Da für diesen Layer keine geeignete Automatisierungstechnologie gefunden wurde, modellierten und implementierten die Teams die dimensionalen Objekte (Dimensions, Facts) manuell.

Vorgehen und Herausforderungen

Die Einführung einer SAP-Retail-Lösung ist keine geringe Herausforderung. Im Rahmen der Einführung entschied sich C&A, parallel eine DWH-Plattform aufzubauen, um bereits zum Start des Pilotbetriebs ein funktionierendes Reporting zur

Verfügung zu haben. Grundsätzlich wurde dieses Ziel zum März 2022 erreicht. Weitere Anpassungen am System führten aber zu zeitgleichem Anpassungsbedarf am DWH. Speziell die nachträgliche Änderung fachlicher Schlüssel machte ein Überarbeiten des Systems notwendig. Hier machte sich die Einführung einer „Persistent Staging Area“ (PSA) bezahlt, mittels derer das konzeptionelle Datenmodell im Raw Vault auch später, inklusive einer kompletten Datenhistorie, nachgeladen werden konnte. Das „Refactoring“ konnte so deutlich vereinfacht werden.

Bereits in dieser frühen Phase zeigte sich, dass die Modellierung des Raw Vault auf Basis eines konzeptionellen Modells eine Entkopplung vom jeweiligen Quellsystem (SAP) ermöglichte. Grundsätzlich orientierten sich die Teams an den SAP-Entitäten und bildeten diese als Hubs ab. Allerdings wurden dann spezielle Satelliten für einzelne fachliche Entitäten gebildet und so schon eine Unterscheidung vorgenommen. Als Beispiele sind hier „Document“ oder „Business Partner“ zu nennen, deren unterschiedliche Ausprägungen dann auch entsprechend in Satelliten abgelegt wurden.

Fachliche Modellierung mit dem Datavault Builder

Natürlich musste ein Toolset ausgewählt werden, das dieses Vorgehen optimal unterstützt und ohne Medienbrüche auskommt. Der Datavault Builder hat sich hier als geeignetes Werkzeug erwiesen. In der Modellierungskomponente lassen sich Entitäten und deren Beziehungen modellieren, ohne einen konkreten Bezug auf ein Quellsystem darstellen zu müssen. Diese einem Entity Relationship Diagram sehr ähnliche Darstellung diente bereits in der ersten Entwicklungsphase als Diskussionsgrundlage mit den fachlichen Anforderern. Sobald sich das Modell gefestigt hatte, konnte in einer zweiten Phase ein Mapping der Quelldaten in das Modell erfolgen.

Im Zusammenhang mit SAP als Datenquelle stellte sich immer wieder die Herausforderung, dass fachliche Schlüssel nicht immer sofort klar erkennbar waren. Die BW-Extraktoren arbeiten schwerpunktmäßig mit technischen SAP-Systemschlüsseln (zum Beispiel DBID), die außerhalb der technischen SAP-Sicht keine Relevanz aus fachlicher Sicht haben. Das integrierte Data-Vault-Modell ist nur sinnvoll über den fachlichen Schlüssel (Business Key) aufzubauen [LiO 15]. Da dieser nicht für alle BW-Extraktoren zur Verfügung stand, galt es, in einem vorverarbeitenden Schritt den fachlichen Schlüssel hinzuzufügen. Das geschah innerhalb der Views des Staging Preparation Layer. So wurde sichergestellt, dass der Data Vault anhand von fachlichen Schlüsseln modelliert werden konnte.

Veröffentlichung der Daten

Der Business Vault ist grundsätzlich dafür da, die fachliche Sicht auf die Daten zu repräsentieren, und ist als Ergänzung zum Raw Vault zu sehen.

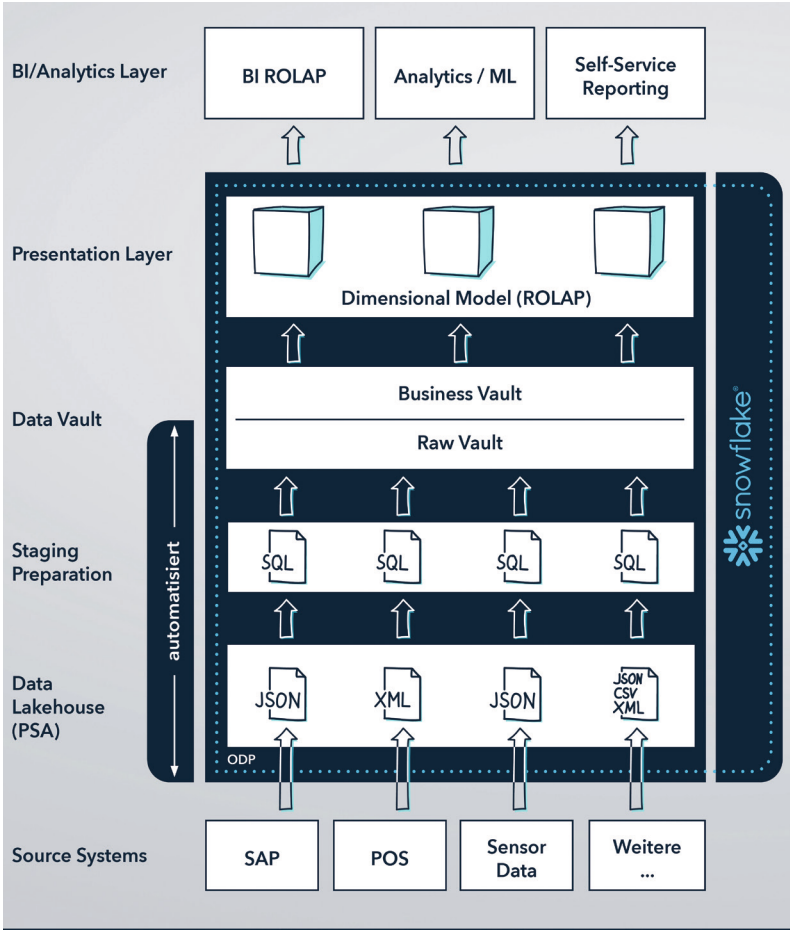

Abb. 1: Layer-Konzept des DWH bei C&A

Kurz gesagt, der Business Vault repräsentiert eine zu den Anforderungen passende Struktur und liefert zudem alle Transformationen, um die jeweiligen Anforderungen zu erfüllen. Das finale Modell orientiert sich daher häufig an der dimensionalen Modellierung, speziell wenn der Abnehmer ein BI-Tool wie MicroStrategy ist, und die Entscheidung der technischen Realisierung (Views) wird dem Anwender ganz im Sinne einer Automatisierung durch den Datavault Builder abgenommen.

Mit dem Einsatz der modernen, analytischen Datenplattform Snowflake ist es nicht mehr in jedem Fall notwendig, Data Marts zu persistieren. Vielmehr reichen oft Views auf das Data-Vault-Modell aus, um entsprechende Strukturen virtuell zur Verfügung zu stellen. Trotzdem zeigen die Erfahrungen im Projekt, dass gerade beim direkten Zugriff des BI-Tools auf die Datenbank eine Persistierung der Business Rules der Virtualisierung vorzuziehen war, um bessere Antwortzeiten zu erreichen.

Grundsätzlich kann man den Standardmechanismus „Business Vault Load“ des Datavault Builder nutzen und schreibt das Ergebnis mit allen Vorteilen der Data-Vault-Modellierung in den Business Vault zurück. C&A hat sich allerdings dafür entschieden, mittels einer eigenen über den Datavault Builder gesteuerten Routine die Business Rules direkt als Snapshot zu persistieren.

Durch die Wahl von SAP BW als Datenquelle ist der Großteil der Geschäftslogik bereits dargestellt

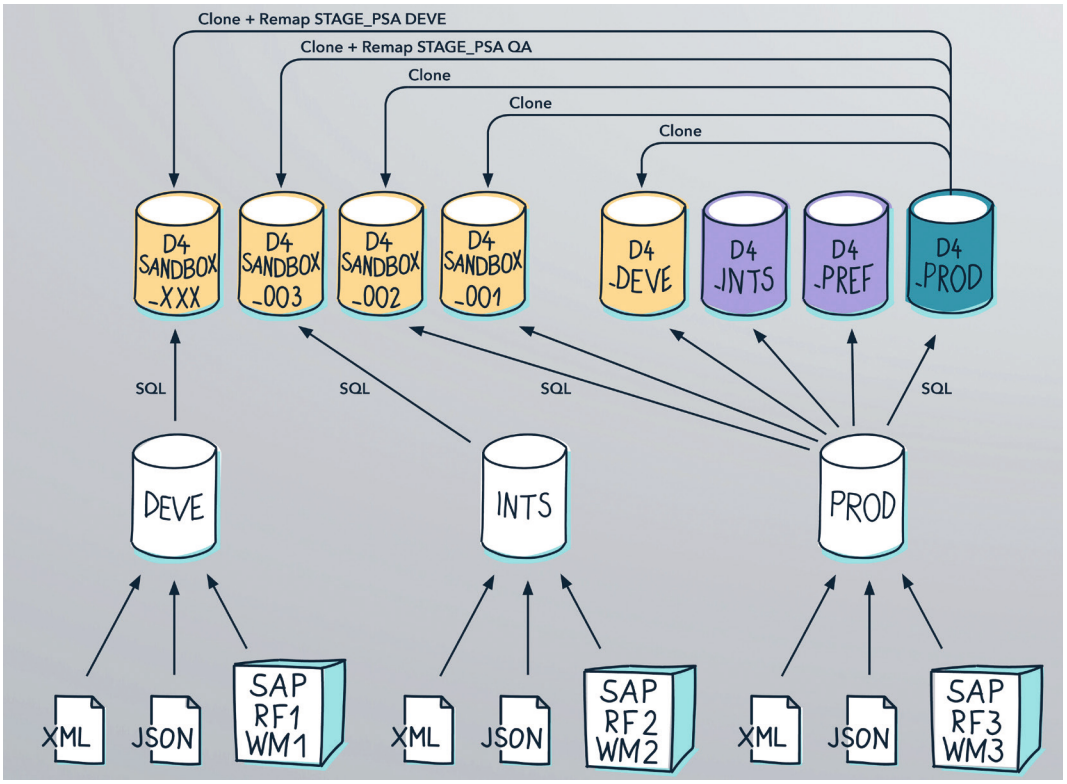

Abb. 2: Zuordnung der Data-Lake-Stufen zu den DWH-Umgebungen und -Sandboxes

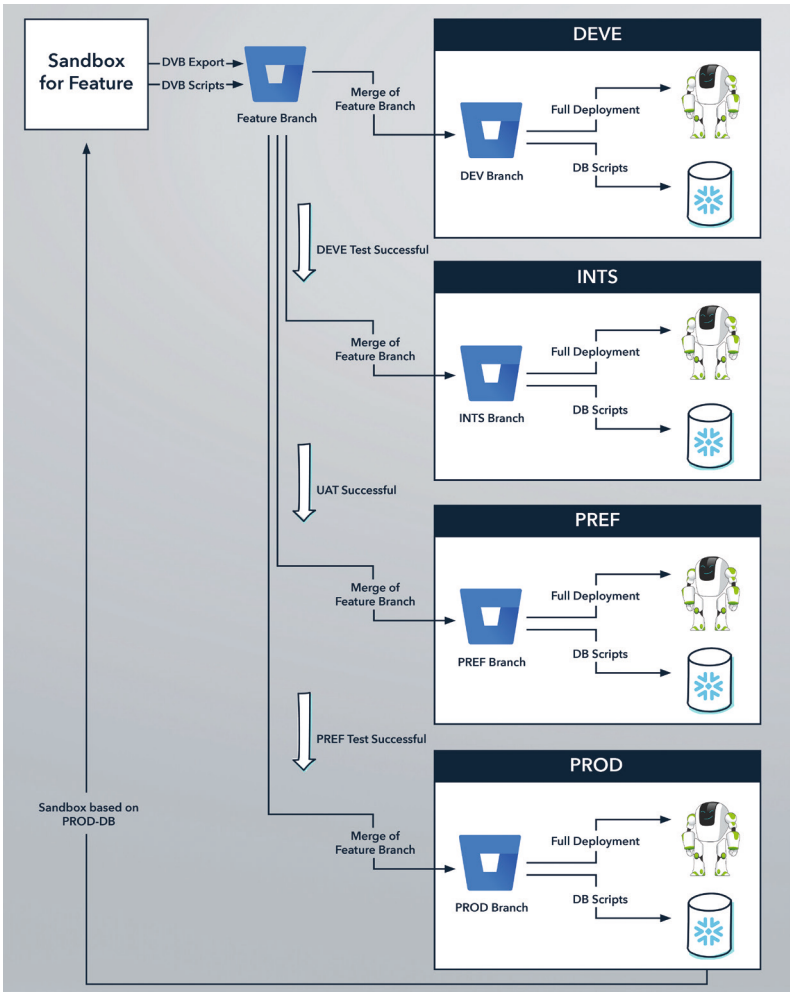

Abb. 3: Deployment-Prozess über die unterschiedlichen Umgebungen von C&A

und musste nur noch in das DWH überführt werden. Damit beschränkte sich der Aufbau des Business Vault im Wesentlichen auf die Strukturierung der Daten für die Abnehmer. In wenigen Ausnahmefällen musste vom Prinzip „Möglichst keine Abbildung von zusätzlicher Geschäftslogik im DWH" aber abgewichen werden. Ein prominenter Fall ist die Bestandsberechnung („Stock Calculation"). Hier werden von Seiten des SAP-ERP lediglich Schnittstellen für Initialbestände und Bestandsbewegungen angeboten. Die Bestände auf Tagesebene sind auf dieser Basis im DWH zu berechnen – inklusive Berücksichtigung von Bestandsbewertungen.

Bereitstellung von Development Sandboxes

Eine weitere Herausforderung bestand in der klassischen mehrstufigen Softwareentwicklung. Das Projekt startete mit der Anbindung des SAP-Entwicklungssystems an die DWH-Entwicklungsumgebung. Es ließ sich jedoch schnell feststellen, dass die Datenabdeckung, zum Beispiel hinsichtlich Verknüpfbarkeit der Geschäftsvorfälle, für eine BI-Implementierung nicht ausreichend war. C&A ist dieser Herausforderung begegnet, indem das produktive Data Lakehouse (PSA) mit den SAP-Produktionsdaten bereits auf der DWH-Entwicklungsstufe zur Verfügung gestellt wird. Das bedeutet: Die Entwicklung des DWH wird bereits auf produktiven SAP-Daten durchgeführt.

Dieses Vorgehen wird optimal durch die eingesetzten Komponenten Snowflake, Datavault Builder und Git unterstützt. C&A setzt drei agile Teams für die Weiterentwicklung des DWH ein. Grundsätzlich ist es möglich, mit dem Datavault Builder mit mehreren Entwicklern beziehungsweise Entwicklerteams auf derselben Umgebung zu arbeiten. Allerdings ist dann die Gefahr groß, dass sich die

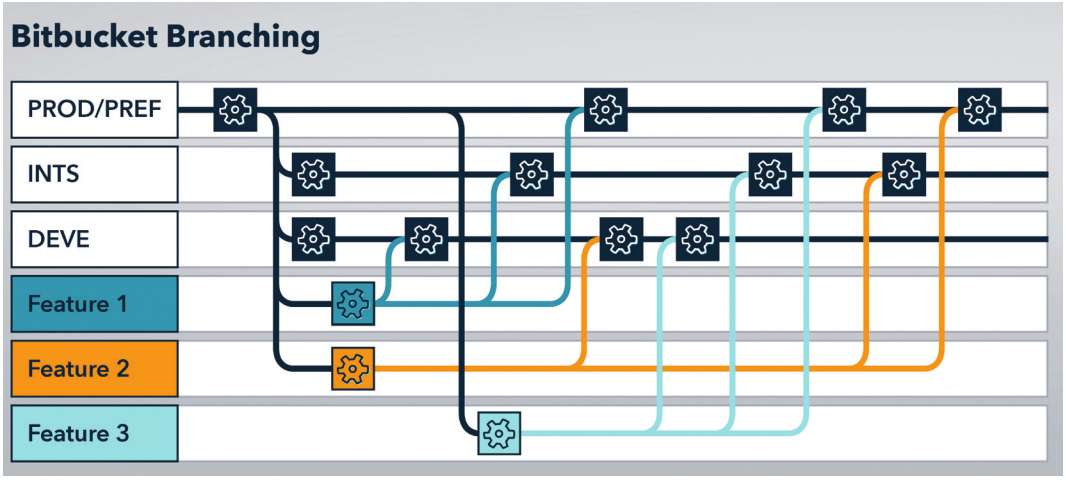

Abb. 4: Implementierung der Features über verschiedene Branches

einzelnen Teams behindern oder sogar Entwicklungsstände überschreiben. Aus diesem Grund ist die klare Empfehlung, Themen- („Features“) oder Team-spezifische Umgebungen („Sandboxes“) aufzusetzen, um hier ein unabhängiges Arbeiten zu ermöglichen.

Auf Seiten der Datenbank wird das sehr einfach gelöst: Durch die Cloning-Funktion der Snowflake wird auf Knopfdruck eine neue Umgebung, in der Regel auf Basis der Produktion, zur Verfügung gestellt (Abbildung 2). Diese neue Umgebung kann ein Team dann nutzen, um das nächste Feature unter realistischen Datenbedingungen zu entwickeln. Da alle Metadaten des Datavault Builder bereits in der Datenbank bzw. in den vom Datavault Builder entwickelten Artefakten enthalten sind, muss sich dieser nur mit der Datenbank verbinden und hat automatisch den korrekten Entwicklungsstand. Nach Abschluss des Features werden die Metadaten des Datavault Builder exportiert und in den Entwicklungs-Branch des Git eingespielt. Da alle Objekte lesbar als JSON abgespeichert werden, können sie leicht zusammengeführt werden. Auch sporadisch auftretende Merge-Konflikte können so gelöst werden. Um aber diese Konflikte weitestgehend zu vermeiden, werden alle Features so geschnitten, dass es keine Überschneidungen zwischen den Teams gibt. So werden alle Features auf der jeweiligen Team-eigenen Sandbox entwickelt und anschließend in den zentralen Entwicklungs-Branch zurückgespielt (Abbildung 3).

Durch die konsequente Trennung der Weiterentwicklung in Features werden alle einzelnen Entwicklungsschritte genau dokumentiert. Da die Features unabhängig voneinander entwickelt werden, wird jedes einzelne Arbeitspaket nach Fertigstellung direkt eingespielt. Einzelne Features überholten so schon früher gestartete Entwicklungen (Abbildung 4).

Fazit

Die Ziele des Greenfield-DWH-Projekts waren: (1) flexibles, performantes Reporting, (2) konsistente Sicht auf die SAP-ERP-Landschaft sowie (3) Unterstützung des agilen Ansatzes durch kurze iterative Umsetzungszyklen.

1. Wurde durch den Einsatz einer skalierbaren Cloud-Technologie wie Snowflake auf AWS erreicht.
2. Wurde durch die Verwendung der SAP-BW-Extraktoren erreicht – welche die Sicht auf die SAP-Daten weitgehend passend für Auswertungszwecke zur Verfügung stellen. Hierdurch lassen sich Aufwandsersparnisse erzielen. Aufwände entstehen jedoch für die Anwendungsbereiche, in denen keine SAP-BW-Extraktoren zur Verfügung stehen beziehungsweise fehlende Attribute durch individuelle Schnittstellen abgebildet werden müssen (Customizing).
3. Für kurze Umsetzungszyklen hat sich der Data-Lakehouse-Ansatz in Kombination mit Data Vault als vorteilhaft erwiesen. Hierdurch kann in einem neuen Sprint zur Erweiterung des Datenmodells sehr rasch auf bestehende Dateninhalte im Data Lake zugegriffen werden. Auch das Refactoring beziehungsweise Auflösen technischer Schulden kann durch Modellanpassungen und Nachladen einfach bewerkstelligt werden.

Ohne eine cloudbasierte Plattform wäre ein breit angelegter Data-Lakehouse-Ansatz allerdings nicht wirtschaftlich möglich. Ein wichtiger Schlüssel für die Agilisierung ist zudem der Einsatz von Automatisierung für die Schnittstellen im Zusammenhang mit Data Vault über den Datavault Builder.

Der Gedanke des Einsatzes von Developer Sandboxes erfordert Umdenken und Übung. Die Idee überzeugt – insgesamt ist der Ansatz im Projekt aber noch neu und der Nutzen zu evaluieren.

Literatur

[Li015] Linsted, D. / Olschimke, M.: Building a Scalable Data Warehouse with Data Vault 2.0. Morgan Kaufman 2015

[KiR13] Kimball, R. / Ross, M.: The Data Warehouse Toolkit – The Definitive Guide to Dimensional Modeling. Wiley 2013