DWH Modernisierung mit Data-Lake, Lab und Governance

Fabian Hardt
OPITZ CONSULTING Deutschland GmbH
Gummersbach

Schlüsselworte:

DWH Modernisierung, Referenzarchitekturen, Data Lake, Data Lab, DWH Offloading.

Einleitung

Inmitten von Digitalisierung und Industrie 4.0 haben sich die Anforderungen für die Speicherung und die anschließende Analyse an klassische, seit vielen Jahren etablierte DWH-Systeme maßgeblich geändert. Bleibt der klassische Ansatz das Mittel der Wahl oder sind Unternehmen gezwungen, die bestehende DWH-Architektur zu modernisieren oder langfristig sogar zu substituieren?
Bei dieser Entscheidung kann es hilfreich sein, sich die Möglichkeiten zur Modernisierung einmal genauer anzusehen. Wie zum Beispiel sollte eine Modernisierung aus technischer Sicht aufgebaut werden, damit die stetig steigenden Anforderungen weiterhin angemessen erfüllt werden können?
Wir möchten in diesem Vortrag einige Architekturszenarien aufzeigen und ihre Vor- und Nachteile einander gegenüberstellen.
Dabei wird deutlich, welche verschiedenen Möglichkeiten es bei der Modernisierung gibt und welche Rolle Data Lakes, Data Labs und die Data Governance dabei spielen. Welche Bedeutung haben diese Schlagwörter und wie sind sie miteinander verwoben? Warum sind viele Big-Data-Technologien unweigerlich mit ihnen verknüpft? Und in welcher Konstellation kann ein Data Lake zu Kosteneinsparungen in der bestehenden Infrastruktur führen?
Des Weiteren berichten wir von einem DWH Offloading Szenario anhand eines Praxisbeispiels und vergleichen hierbei den Einsatz von Oracle- vs. Open Source Technologien.

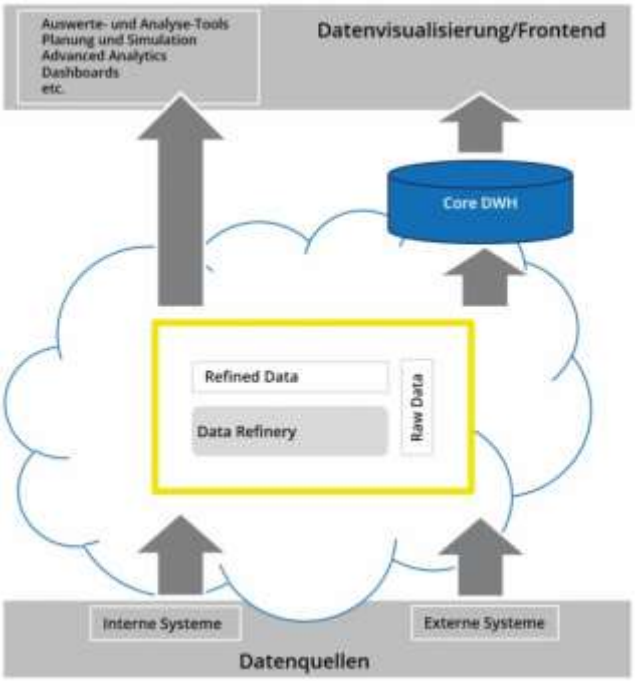

Abbildung 1 Sequentielle Architektur

Unsere DWH Modernisierungsvorschläge - Architektursteckbriefe

Als Erstes möchten wir auf eine **sequentielle Architektur** eingehen. Hierbei wird die klassische Staging Area eines DWHs durch einen Data Lake ersetzt. Dies bietet den Vorteil, dass die komplexen Kennzahlberechnungslogiken und bereits qualitätsgesicherten Ladestrecken beibehalten werden. Es müssen neue Ladestrecken aus den Quellsystemen in den Data Lake hinein

gebaut werden, sowie die ehemaligen Staging-Tabellen abgelöst werden. Dies könnte im Hadoop Umfeld mittels Hive passieren, sodass die Ladestrecken die Daten in den Core hinein aus strukturierten Tabellen auslesen können.

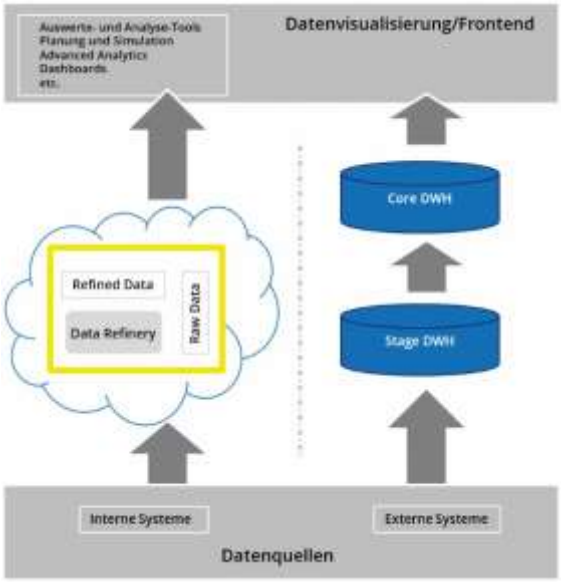

Abbildung 2 Parallele Architektur

Die **parallele Architektur** ist eine naheliegende Architekturform, die gänzlich getrennt vom DWH aufgebaut wird. Es gibt keine Datentransfers zwischen den Systemen.
Diese Architekturform erleichtert die Organisation innerhalb des Unternehmens, da die Systeme durch zwei komplett getrennte Abteilungen aufgebaut und gewartet werden können.
Ein großer Nachteil liegt darin, dass die gewonnenen Insights des DWHs nicht im Data Lake genutzt werden können. Diese Erkenntnisse müssen neu gewonnen werden. Alles in allem eignet sich diese Architektur nur für explorative Zwecke, um Daten zu analysieren, die bislang noch nicht ausgewertet werden. Möchte man komplexere Zusammenhänge erkennen, die sich über alle Unternehmensbereiche erstrecken, wird ein Zusammenspiel mit dem DWH immer wichtiger.

In einer **Offloading Architektur** läuft der Prozess genau in umgekehrter Reihenfolge ab. Auch hier haben die qualitätsgesicherten Ladestrecken des DWHs weiter Bestand, der Data Lake ist hier jedoch Ziel der Daten und nicht die Quelle. Die aufbereiteten Daten, wie z.B. Kennzahlen werden im Data Lake zur Verfügung gestellt, sodass diese von Data Scientists genutzt werden können.
Oft wird diese Architekturvariante aus Kostengründen gewählt. Insbesondere alte Daten (typischerweise Daten der Faktentabellen) können kostengünstig in einen Data Lake ausgelagert werden. Somit können hohe Lizensierungskosten für eine Enterprise Datenbank eingespart werden.

Ein klarer Nachteil dieser Architektur liegt jedoch auf der Hand: Unstrukturierte Daten der DWH Quellsysteme werden nicht in den Data Lake aufgenommen, da sie nicht durch das DWH geschleust werden können.

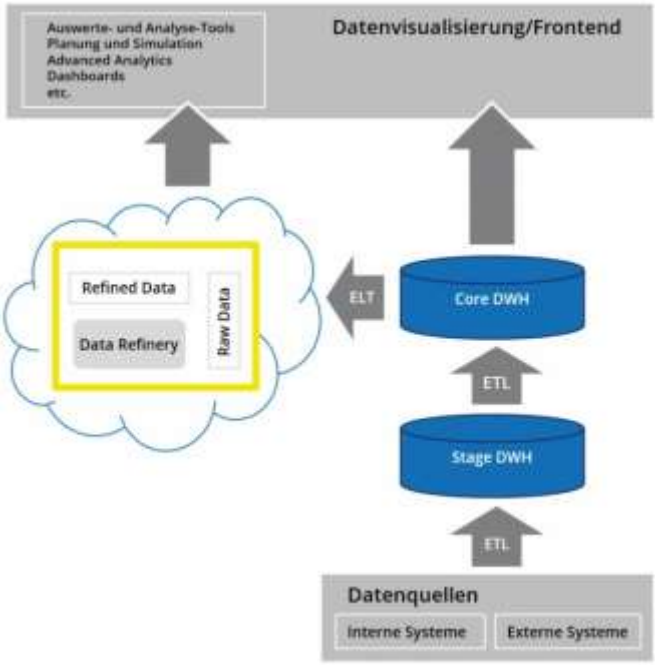

Abbildung 3 Offloading Architektur

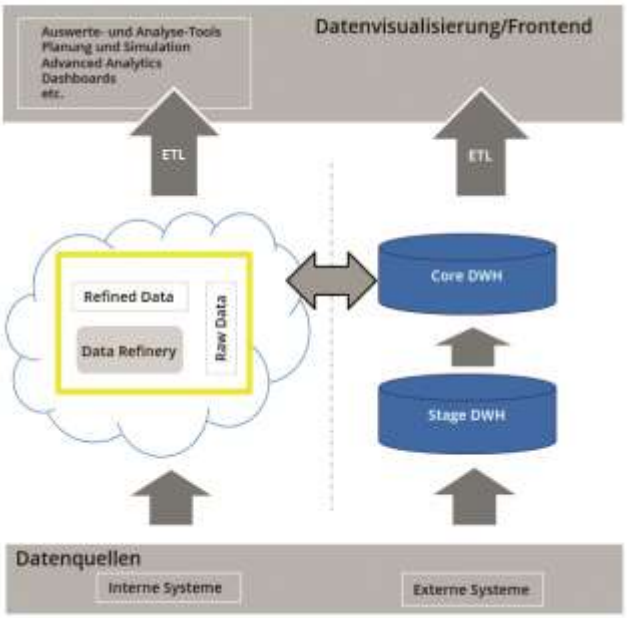

Abbildung 4 Hybride Architektur

Kommen wir zur dritten und umfangreichsten Modernisierung, welche auch gleichzeitig die größte Flexibilität bietet. Wir sprechen hierbei von einer **hybriden Architektur**. Die Quellsysteme sind sowohl am Data Lake, als auch am DWH angebunden. Außerdem findet ein Offloading der qualitätsgesicherte Daten in den Data Lake statt. Dies gilt sowohl für Daten aus dem Core-Layer, welcher Unternehmensknowhow aus diversen Fachbereichen enthält, als auch die Daten der Data Marts, insbesondere die berechneten Inhalte der Faktentabellen.
Der wohl größte Nachteil dieses Szenarios besteht in den hohen Implementierungs- und Wartungskosten, die hierfür nötig sind. Bei Änderungen an den Quellsystemen müssen sowohl die Data Lake Ladestrecken, als auch die Ladestrecken ins DWH hinein angepasst werden. Beide „Datenlager“ müssen durch die IT ausreichend betreut werden. Dies bedeutet in der Regel personellen Mehraufwand, da zusätzliches Big Data Know-How benötigt wird, welches die vorhandenen DWH Entwickler nicht zusätzlich zu ihren normalen Aufgaben leisten können.

Bedeutung von Metadaten im Data Lake

Die Schemalosigkeit von alternativen Speichermöglichkeiten bringt jedoch nicht nur Vorteile mit sich. Die neu gewonnene Freiheit kann schnell im Chaos enden, wenn man kein stringentes Konzept zur Metadatenverwaltung hat.
Um zu verstehen, warum Metadaten-Management so wichtig ist, müssen zunächst die drei Arten von Metadaten erläutert werden.

1. **Technische Metadaten**

Diese Metadaten gehören typischerweise eng zu den Daten. Beispiele wären Informationen zum Quellsystem, zu Tabellen oder deren Spalten, sowie Datumsstempel, also wann wurden die Daten geladen, und viele weitere.

2. **Fachliche Metadaten**

Für den Konsumenten der Daten sind dies die wohl wichtigsten Metadaten. Sie geben in gut indizierbaren Stichworten an, was sich fachlich hinter diesen Daten verbirgt. Bezogen auf Aggregationstabellen könnten dies zum Beispiel Informationen sein, aus welchen Quelltabellen diese berechnet wurden, oder sogar eine Kurzfassung der Berechnungsvorschrift.

3. **Operative Metadaten**

Diese Metadaten sind ebenfalls technisch geprägt. Sie stammen allerdings nicht aus der Quelle, sondern werden direkt im Data Lake erhoben. Dies sind zum Beispiel historisierte Daten, welcher Benutzer auf welche Daten zugegriffen hat und ob diese vielleicht sogar geändert wurden. Der Begriff operativ bezieht sich also auf den Betrieb des Data Lakes.

Insbesondere in Hinblick auf die aktuelle Diskussion zur EU-Datenschutz-Grundverordnung spielt die Verwaltung von Metadaten eine besondere Rolle. In den meisten Fällen werden in einem Data Lake auch personenbezogene Daten aufbewahrt, doch gerade diese obliegen besonders strengen Regeln. Hierbei können Metadaten helfen, denn sie zeigen uns den Zweck der Datenspeicherung auf und können ggf. auch die erlaubte Speicherdauer beinhalten. Die EUDSGV fordert „Rechtmäßigkeit, Verarbeitung nach Treu und Glauben,[und] Transparenz“, diese Paradigmen sind also auch beim Aufbau eines Data Reservoirs (Refined Data Bereichs) zu beachten.
Zusammenfassend kann man also sagen, dass Metadaten-Management im Data Lake extrem wichtig ist um den Überblick zu behalten. Nur wenn dies umfassend genug und automatisiert abläuft, wird ein Data Lake zu einem echten Mehrwert im Unternehmen.
Wie die Vorteile von Data Lake, Data Lab und Data Governance in klassische DWH-Konzepte einfließen können, zeigen die folgenden Architekturszenarios zur Modernisierung eines bestehenden DWH.

Das erwartet Sie in unserem Vortrag

In unserem Vortrag bekommen Sie eine kompakte Einführung in Data Lakes, Data Labs und Data Governance, alles im Kontext einer DWH Modernisierung. Außerdem bekommen Sie einen Überblick über von uns ausgearbeitete Referenzarchitekturen rund um dieses Thema.
Als Abschluss des Vortrages erwartet Sie eine Diskussion zum DWH Offloading, anhand einiger Praxisbeispiele. Hierbei setzen wir zum einen auf Enterprise Tools aus dem Hause Oracle, zum anderen auf eine Lösung mit Open Source Tools.

Kontaktadresse:

Fabian Hardt
OPITZ CONSULTING Deutschland GmbH
Kirchstraße, 6
D-51647 Gummersbach

Telefon: +49 (0) 2261-6001 1045
Fax: +49 (0) 2261-6001 4200
E-Mail fabian.hardt@opitz-consulting.com
Internet: www.opitz-consulting.com