



Whitepaper

# Data Warehouse: Drei Lösungsansätze zur Modernisierung



**it-novum GmbH Deutschland**

Hauptsitz Fulda: Edelzeller Straße 44 • 36043 Fulda  
Telefon: +49 661 103 333  
Niederlassungen in Düsseldorf & Dortmund

**it-novum Zweigniederlassung Österreich**

Ausstellungsstraße 50 / Zugang C • 1020 Wien  
Telefon: +43 1 205 774 1041

**it-novum Schweiz GmbH**

Hotelstrasse 1 • 8058 Zürich  
Telefon: +41 44 567 62 07

# Inhalt

1. Einleitung	4
2. Modernisierung mit einem Cloud Data Warehouse	6
2.1 Ausgangslage	6
2.2 Vorteile eines Data Warehouse in der Cloud im Allgemeinen	8
2.3 Herausforderungen bei einer Cloud-Lösung	10
2.4 Lösung: Moderne Data Warehouse-Architektur in der Cloud von Snowflake	12
2.5 Einsatzszenario Energiewirtschaft	14
3. Modernisieren mit Data Vault	16
3.1 Ausgangslage	16
3.2 Data Warehouse-Modellierung mit Data Vault	17
3.3 Vorteile und Herausforderungen von Data Vault	19
3.4 Praxisbeispiel: Unternehmensübernahme	20
3.5 Entscheidung leichter gemacht: Ist Data Vault für Sie geeignet?	20
4. Modernisieren durch pseudonymisierte Personendaten	22
4.1 Ausgangslage	22
4.2 Lösung: Der abweichende Zwilling	23
4.3 Praxisbeispiel für pseudonymisierte Personendaten im Data Warehouse	24
4.4 Ihre Entscheidung leichter gemacht: Verarbeitet Ihre IT die Daten DSGVO-konform?	25
5. So machen Sie Ihr Data Warehouse zukunftsfest	26

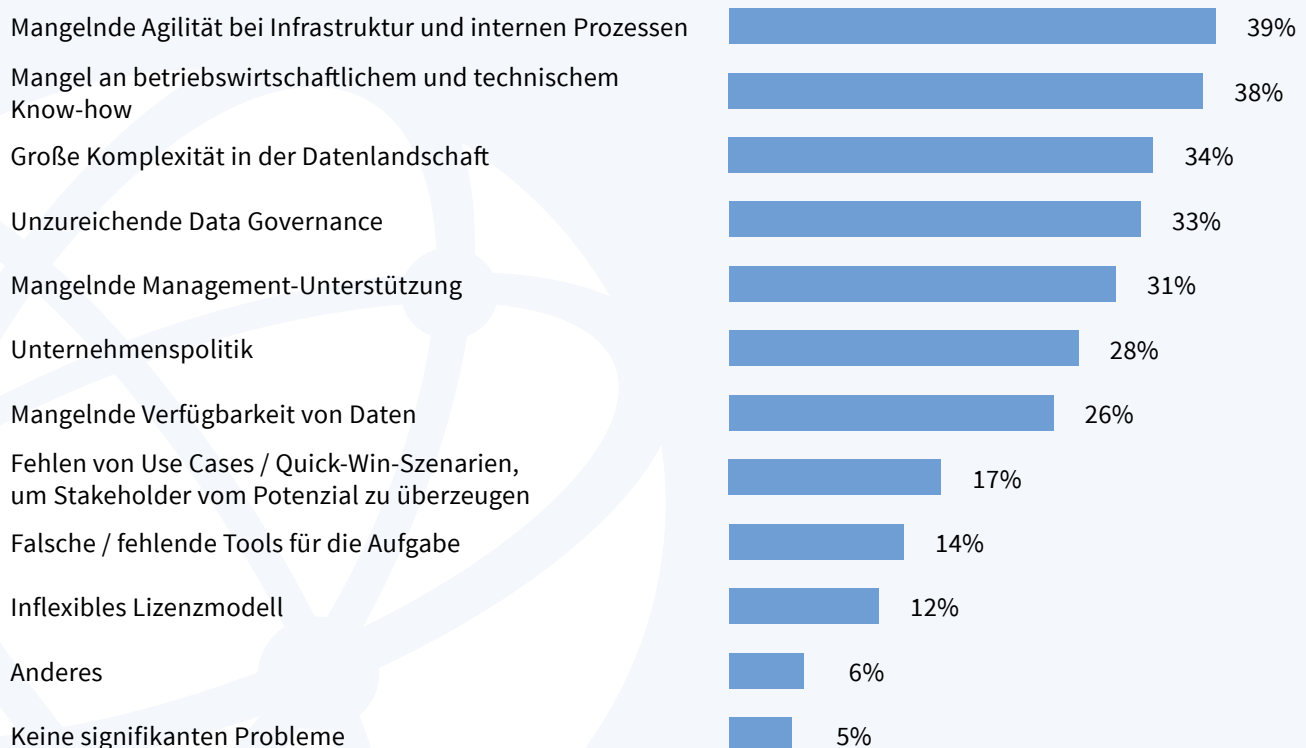


# 1. Einleitung

Im Rahmen der digitalen Transformation steigt der Druck auf die Unternehmen, ihre Daten effizient und zielgerichtet zu nutzen. Traditionelle (on-premise) Data Warehouses (DWH) stoßen dabei oft an ihre Grenzen.

Nach einer BARC-Umfrage steht daher die Modernisierung des Data Warehouse bei vielen Unternehmen oben auf der Prioritätenliste, um der zunehmenden Komplexität der Datenlandschaft entgegenzuwirken und ihre Infrastruktur und internen Prozesse agil halten zu können. Dabei stehen die Verantwortlichen vor folgenden Herausforderungen:

## Was sind die größten Herausforderungen, denen Ihr Unternehmen bei der Modernisierung seiner Data-Warehouse-Umgebung gegenübersteht?



Quelle: BARC Research: Modernizing the Data Warehouse: Challenges and Benefits, 2019

Im vorliegenden Whitepaper stellen wir drei Möglichkeiten vor, wie sich die DWH Modernisierung meistern lässt. Dabei liefern wir Antworten auf verschiedene Fragen die im Kontext der Modernisierung auftauchen:

- wie muss eine moderne DWH-Architektur in der Cloud aufgebaut sein?
- was ist in Hinblick auf Data Vault als vielversprechender Technologie zu beachten?
- wie können Daten pseudonymisiert werden, um DSGVO konform weiterhin alle Usecases umzusetzen?

Ziel ist es, das DWH so zu gestalten, dass es agil, performant und skalierbar wird und durch die Erfüllung gesetzlicher Datenschutzvorgaben auch zukünftige Herausforderungen abdecken kann.

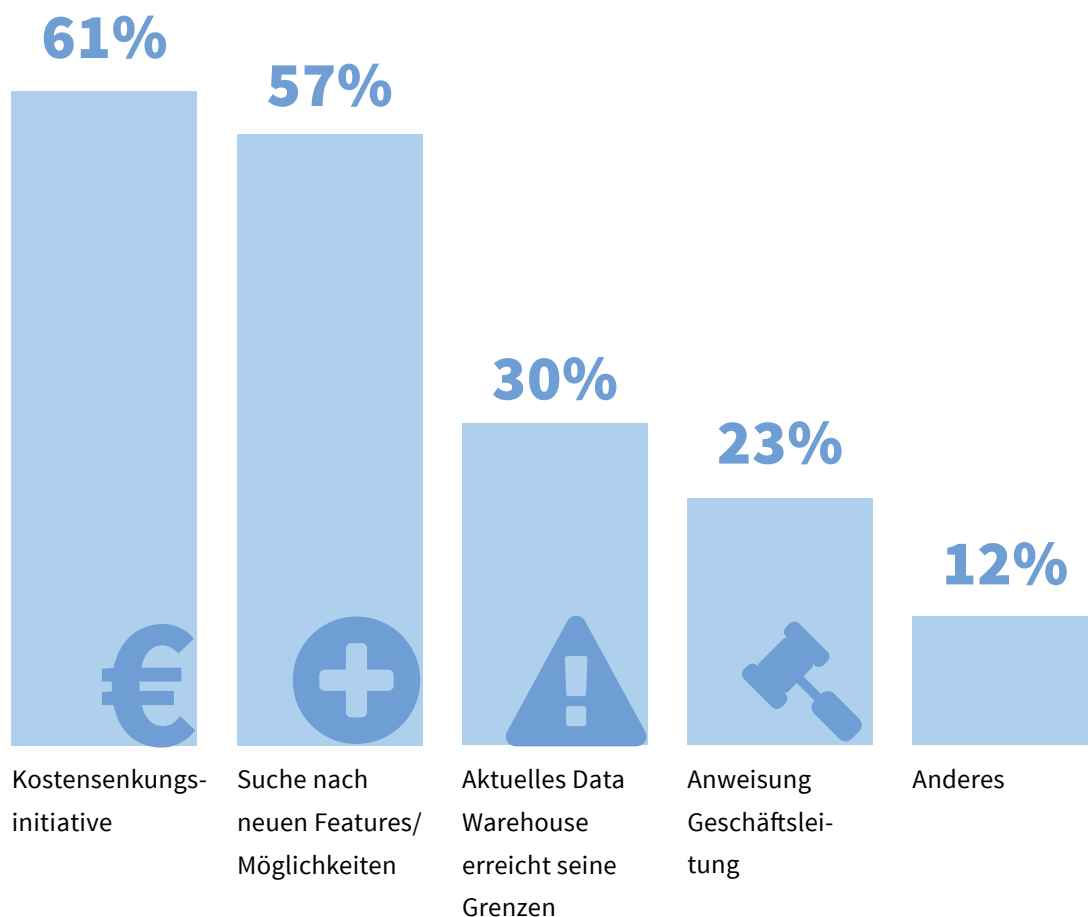
## 2. Modernisierung mit einem Cloud Data Warehouse

### 2.1 Ausgangslage

Die Anforderungen der Fachabteilungen an die Datenverarbeitung werden zunehmend komplexer. Es geht darum, immer heterogenere Daten auswerten zu können, von Informationen aus anderen ERP und Fachsystemen über IoT-Sensordaten und Maschinenformaten bis hin zu Logs, Text- und Videodateien. Zudem sind die unterschiedlichsten internen und externen Datenquellen einzubeziehen, egal, ob es sich dabei um ein ERP-, CRM- oder ein anderes operatives System handelt.

Damit und mit den oft einhergehenden gestiegenen Anforderungen wie Skalierung und Geschwindigkeit stoßen On-Premise-Data Warehouses häufig an ihre Grenzen. Dies belegt die oben erwähnte BARC Studie, in der die Befragten als Hauptgrund für ihre eingeschränkte Handlungsfähigkeit die mangelnde Flexibilität ihrer vor-Ort Infrastruktur beklagen.

Eine Lösungsmöglichkeit ist das Data Warehouse in der Cloud. In einer Datometry Studie nennen Entscheider folgende Gründe für die Migration von On-Premise Data Warehouses in die Cloud:



Quelle: Datometry: Cloud-First Enterprise: The Time Is Now, o.J.

*Datensilos, unstrukturierte und semistrukturierte Daten sowie der eingeschränkte, standortabhängige Zugriff erschweren Unternehmen die nutzbringende Auswertung vorhandener Informationen.*

## Typische Einschränkungen eines On-Premises-DWH



### Datensilos und verschiedenste Datenquellen

- Viele Datenquellen/-stores
- Semi-strukturierte Daten
- Datasharing ist erschwert



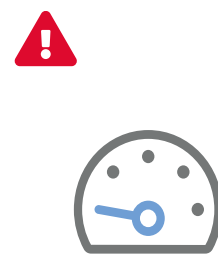
### Skalierungs- und Performance Probleme

- Kann nicht beliebig einfach vergrößert/verkleinert werden
- Keine Trennung zwischen Storage und Compute
- Ressourcen-Konflikte



### Komplexe, teure Infrastruktur

- Teure Software und Hardware
- Überdimensioniert & teils schlechtere Auslastung
- Knowhow für Wartung/Tuning notwendig



### Langsame, begrenzte Entscheidungsunterstützung

- Performance zu Spitzenzeiten nicht immer an Bedarf angepasst
- Teils hohe Kosten für Datenzugriffe

## 2.2 Vorteile eines Data Warehouse in der Cloud im Allgemeinen

Beim Hosting in der Cloud sinkt der für das DWH anfallende administrative Aufwand im Unternehmen. Die Arbeitsbelastung der Datenbankadministratoren sinkt, was die Unternehmens-IT entlastet. Da Cloud-Dienstleister in der Regel zeitnah und nach dem pay-as-you-go-Modell abrechnen, lässt sich der Budgetbedarf exakt an das Nutzungsverhalten ausrichten. Es wird zudem weniger Kapital für eine entsprechende Rechenzentrums-Infrastruktur gebunden. Doch das Wichtigste sind die nahezu unbegrenzte Flexibilität bei der Skalierung und Anpassung von Services an ihre sich ständig ändernden Ziele und Bedürfnisse und die große Schnelligkeit, mit der eine Cloud-Lösung reagieren kann.

Entgegen den nach wie vor ausgeprägten Bedenken in der deutschen Wirtschaft kann die Datenhaltung in der Cloud sogar sicherer sein als On-Premise. Alle großen Cloud-Anbieter sind nach der höchsten Sicherheitsstufe zertifiziert und sorgen zudem für die Pflege und regelmäßige Updates der Infrastruktur in der Cloud – eine Situation, von der viele IT-Abteilungen eher entfernt sind.

### **Datensicherheit und Cloud passen gut zusammen**

Das liegt einerseits daran, dass zertifizierte Cloud-Anbieter besser gegen Risiken geschützt sind als das Rechenzentrum des einzelnen Unternehmens. Besonders die Global Player im Cloud-Bereich haben die höchsten Sicherheitszertifizierungen. Ganze Teams von Sicherheitsspezialisten beim Provider stehen einer einzigen IT-Abteilung im Unternehmen gegenüber, die sich um die gesamte IT-Landschaft ihrer Organisation kümmern muss und deshalb zumeist keine Security-Experten, sondern Generalisten umfasst. Dazu kommt: Wird das DWH in der EU, z.B. in Frankfurt als dem größten Internet Exchange Point Europas, gehostet, dann ist der Datenschutz durch die DSGVO gewährleistet.

*Es ist potenziell unsicherer, im eigenen Data Center zu hosten als bei globalen Cloud-Anbietern.*



Ein weiterer Vorteil besteht darin, dass im Cloud DWH die Datenverarbeitung vom Storage getrennt ist. Gegenüber herkömmlichen DWH zeichnet sich dadurch ein Cloud DWH durch eine sehr hohe und flexible Skalierbarkeit – sowohl bei der Rechenleistung als auch beim Speicherplatz – und durch beschleunigte Abfragen aus.

*Bei der Anbieterauswahl sollte man beachten, dass nicht alle Cloud DWHs das Verarbeiten von unstrukturierten Daten (z.B. Videos) oder das direkte Ausführen von Machine Learning-Anwendungen zulassen.*

## 2.3 Herausforderungen bei einer Cloud-Lösung

### — **Ausfallsicherheit**

Generell lässt sich sagen, dass die am Markt verfügbaren Cloud DWH sehr sicher und verlässlich sind. Mehr noch: Sicherheitsupdates und zentral gesteuerte Patches schützen sicher vor Datenverlust und Systemausfällen – besser zumeist, als das für die klassischen DWH der Rechenzentren gilt.

### — **Multi-Cloud**

Es gibt in Unternehmen einen Trend hin zur Multi-Cloud-Strategie. Indem sie ihre Workloads auf mehrere Clouds verteilen, versprechen sich Unternehmen eine größere Unabhängigkeit vom einzelnen Anbieter. Es ist technisch jedoch anspruchsvoll, ein DWH über verschiedene Clouds auszurollen und dabei Brüche zu vermeiden. Auf der Minusseite stehen zudem ein zusätzlicher Aufwand im Management sowie bei der operativen Steuerung, denn jede Cloud-Plattform hat eigene Besonderheiten.

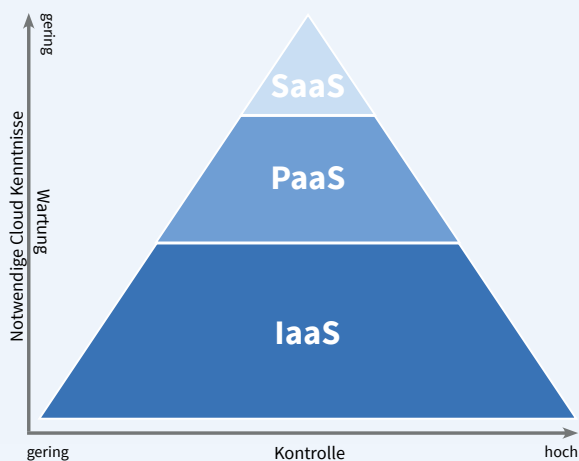
### — **Cloud-Services**

Da sie nicht genügend über die Möglichkeiten und Vorteile rund um die Cloud-as-a-Service-Angebote wissen, verzichten Administratoren häufig auf diese Dienste. Wegen der Notwendigkeit, sich flexibler und agiler aufzustellen, der fortschreitenden Vernetzung über Standort- und Unternehmensgrenzen hinweg sowie aufgrund des zunehmenden Drucks zur umfassenden Datennutzung wird mittel- und langfristig jedoch kein Weg an der Cloud vorbeigehen.

### XaaS: Alles-als-ein-Service

Unternehmen können – je nachdem welche Lösung den größten Benefit bringt – ihre gesamte IT-Landschaft, aber auch nur einzelne Workloads, Applikationen oder Storages, in die Cloud auslagern.

## Die Besonderheiten von SaaS (Software-as-a-Service), PaaS (Platform-as-a-Service) und IaaS (Infrastructure-as-a-Service)



- Komplettlösung: Anwender ist ausschließlich Nutzer der Software
- Anpassungen sind nur eingeschränkt möglich
- Sehr geringe Einstiegshürden

- Kontrolle ab Plattform-Level (Konfiguration notwendig)
- Meist Einsatz für Entwicklertools
- Keine Kontrolle über Hardware
- Mischung aus IaaS und SaaS

- Bereitstellung von Infrastruktur (virtuell, bare-metal oder Container)
- Vollkontrolle ab Hardware durch Anwender (tiefe Cloud-Kenntnisse notwendig)
- Größte Freiheit zur Abbildung individueller Anforderungen

## 2.4 Lösung: Moderne Data Warehouse-Architektur in der Cloud von Snowflake

Es reicht nicht aus, ein vorhandenes Data Warehouse in die Cloud zu verschieben und dort wie gehabt weiter zu betreiben. Deshalb hat der Cloud-Anbieter Snowflake einen besonderen Weg gewählt: die Datenbankarchitektur für das Cloud DWH wurde von Grund auf neu entwickelt und auf die Besonderheiten der Cloud zugeschnitten.

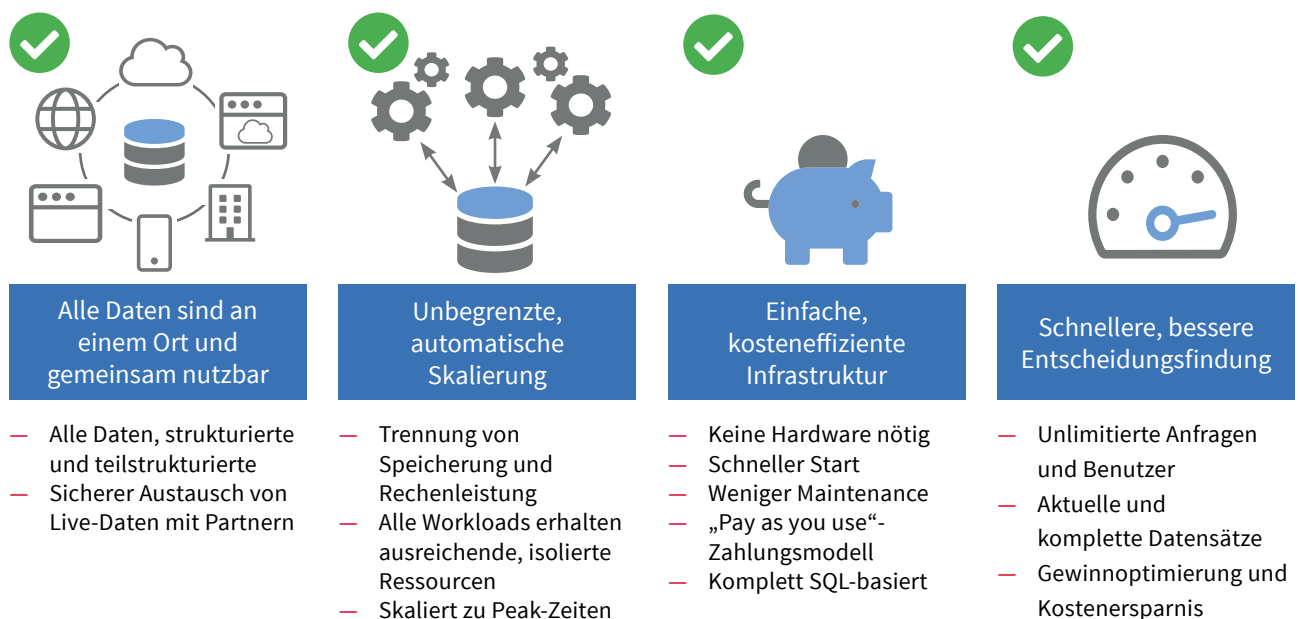
Snowflake nutzt als momentan einziger Cloud DWH-Anbieter keine eigene Public Cloud, sondern stellt das DWH wahlweise bei AWS, Google Cloud Platform oder Microsoft Azure zur Verfügung. Selbst eine horizontale Skalierung über zwei oder alle drei Anbieter ist möglich (Stichwort Multi-Cloud).

*Bei einem lokalen Cloud DWH-Anbieter ist der Nutzer an den Standort im eigenen Land gebunden. Im Gegensatz dazu kann man sich bei der Snowflake-Lösung entscheiden, im Inland zu hosten oder – aus welchen Gründen auch immer – auf einen anderen Staat oder sogar Kontinent auszuweichen.*

Im Ergebnis der konsequenten Trennung von Storage und Datenverarbeitung ist Snowflake höchst performant und flexibel skalierbar. Wenn es rechenintensive ETL-Prozesse erfordern, werden die Computing-Ressourcen automatisch und im notwendigen Umfang hochskaliert. Hier lauert jedoch auch ein teurer Stolperstein: da der Mechanismus automatisch erfolgt, kann die Leistungssteigerung schnell kostspielig werden.

Weitere interessante Funktionen von Snowflake sind das Cachen von Abfrageergebnissen, eine Timetravel-Funktion (die SQL-Funktion „UNDROP TABLE“) und die Möglichkeit, eigene Datensets ohne Replikation mit Kunden oder Geschäftspartnern zu teilen.

## Snowflake-Datenplattform ist angepasst an die Bedürfnisse der Cloud



Snowflake selbst bietet keine eigenen ETL-Applikationen an, sondern überlässt diesen Prozessschritt den Anwendern. Hier bieten BI-Tool-Anbieter meist native Treiber oder Konnektoren an, damit sich ihre ETL-Lösungen mit Snowflake verwenden lassen. Eine empfehlenswerte Datenintegrations- und Analytics-Lösung ist hier Pentaho von Hitachi Vantara, die eine dedizierte und bewährte Konnektivität zu Snowflake bietet.

## 2.5 Einsatzszenario Energiewirtschaft

Das folgende Beispiel zeigt, wo Cloud Data Warehouses ihre Stärken ausspielen. Wie können Energieerzeuger eigene Kapazitäten mit der Einspeisemenge Dritter und dem Energiebedarf in Übereinstimmung bringen? Indem sie gleichzeitig Bedarfsschwankungen und Einspeisemengen von grünem Strom durch die Auswertung historischer Daten genauestens vorhersagen.

Manche Energieunternehmen gehen inzwischen sogar noch einen Schritt weiter und lassen sich mithilfe von Predictive Tools künftige Entwicklungen prognostizieren. So weiß das System aufgrund meteorologischer Daten, dass sich eine bestimmte Wetterlage einstellen wird. Gleichzeitig hat es gelernt, wie hoch Energiebedarf und Einspeisung unter diesen Umständen sind, und berücksichtigt sogar noch, dass sich beispielsweise ein Industrieverbraucher aufgrund von Feiertagen etwas anders verhält als sonst. Diese Prognose mündet in einer Handlungsempfehlung an den Stromlieferanten, etwa, wie er seine Turbinen optimal steuern sollte.

Ein leistungsfähiges Werkzeug für diese komplexe Datenverarbeitung ist das Cloud DWH von Snowflake. Um ein Maximum an Erkenntnissen aus den vorhandenen riesigen Datenmengen, etwa zu kunden- und tageszeitabhängigen Energieverbräuchen, zu ziehen, verknüpft man diese mit Informationen externer Quellen, beispielsweise mit Wetter- oder Geodaten. Diese liegen sehr wahrscheinlich in semistrukturierter Form vor und umfassen ebenfalls eine gigantische Anzahl an Datensätzen.

Das Hochladen dieser Datenbanken, zum Beispiel aus der Stage in Azure, in die Snowflake-Plattform geht schnell und nimmt – je nach Umfang – nur Sekunden bis Minuten in Anspruch.

Ein weiterer Pluspunkt: die semistrukturierten Daten (bei den Wetterinformationen umfassen sie beispielsweise Werte wie Temperaturen, Windgeschwindigkeiten und Niederschlagsmengen und die Descriptions) lassen sich über eine Snowflake-Notation sofort abfragen, ohne sie vorab in eine strukturierte Tabelle zu überführen. Auch ohne ein vorher implementiertes relationales Datenmodell erlaubt Snowflake selbst komplexere Abfragen und bietet verschiedene Möglichkeiten, semistrukturierte Daten zu verarbeiten.



### **Data Warehouse in der Cloud: so macht es Sinn**

Um entscheiden zu können, ob Ihnen ein Data Warehouse in der Cloud Vorteile bringt, sollten Sie sich folgende Fragen stellen:

- Ist die Einführung eines Cloud DWH mit derzeit noch geringen Datenmengen eine gute Gelegenheit, sich mit der Thematik auseinanderzusetzen und von der Einfachheit eines skalierbaren, nutzerfreundlichen Systems zu profitieren?
- Ist zeitnah geplant, ein DWH in Ihrem Unternehmen einzuführen?
- Nutzen Sie bereits ein On-Premise-DWH, sind jedoch unzufrieden wegen der unzureichenden Skalierbarkeit und langen Abfragezeiten?

Wenn Sie eine dieser Fragen bejahen, dann sollten Sie erwägen, eine cloud-basierte Datenplattform einzuführen.

Die Unterschiede bei den führenden Cloud-Anbietern hinsichtlich Funktionsumfang und Performance sind eher marginal. Wenn Sie allerdings besonderen Wert darauf legen,

- das DWH mit relativ wenig Aufwand nach Ihren individuellen Erfordernissen konfigurieren zu können,
- unabhängig von einem Cloud-Anbieter zu sein,
- das DWH sogar über mehrere Clouds zu skalieren,

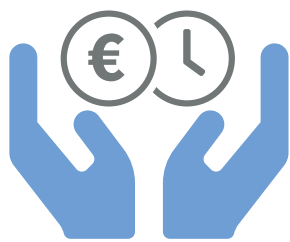
dann wäre das Snowflake Cloud DWH Ihre erste Wahl.

## 3. Modernisieren mit Data Vault

### 3.1 Ausgangslage

Je mehr relevante Daten in eine Analyse einfließen, umso aussagekräftiger ist das Ergebnis. Unternehmen haben daher ein Interesse, immer neue, zusätzliche Datenquellen in ihre Auswertung zu integrieren. Doch mit zunehmender Anzahl steigt auch die Komplexität, zumal es dabei auch eine große Anzahl an Abhängigkeiten und Auswirkungen zu beachten gibt.

Stattdessen sollte bei der Datenmodellierung im DWH angesetzt werden. Das gelingt am besten mithilfe des Data Vault-Konzepts, das Erweiterbarkeit, Historisierung und Zeitbezüge ermöglicht. Häufig ist es damit sogar möglich, die Änderungen auf der vorhandenen DWH-Architektur auszuführen.

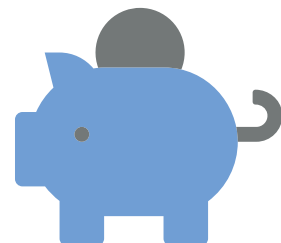


**Up to 88%**  
**savings!**



**Data Vault 2.0 can help your  
IT and Analytics teams  
deliver up to 84% faster,**

**using only 13% of your  
resources at a fraction of  
your budget.**



*Quelle: Data Vault Alliance 2020, [datavaultalliance.com/about/what-is-datavault](https://datavaultalliance.com/about/what-is-datavault)*

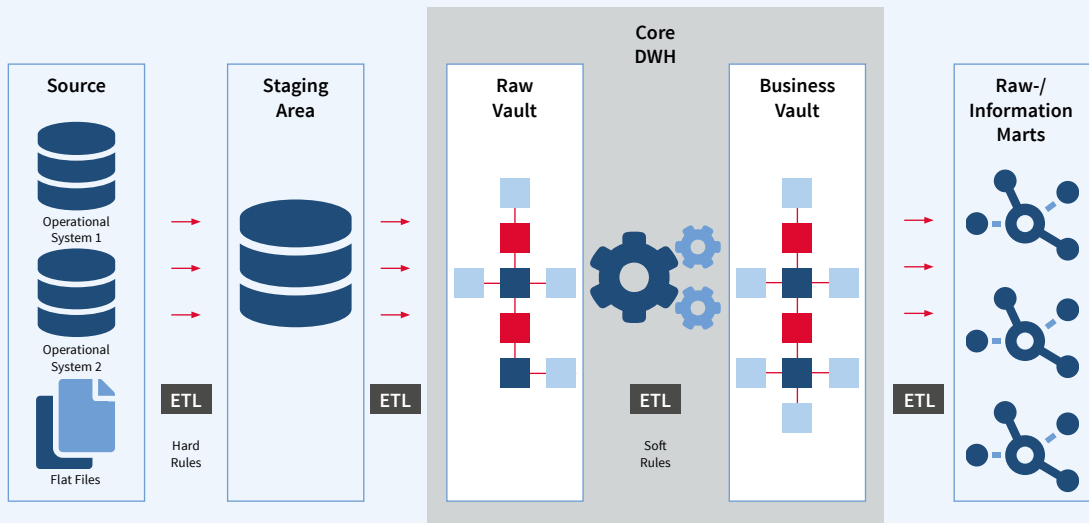
## 3.2 Data Warehouse-Modellierung mit Data Vault

Data Vault ordnet alle zum Objekt gehörenden Informationen drei verschiedenen Kategorien zu:

- Hubs (Identifikation der Entitäten – Geschäftsobjekte),
- Links (Beziehungen zwischen Hubs) und
- Satelliten (beschreibende Informationen bzw. Attribute für Hubs und Links).

Diese Trennung der Informationen in drei Komponenten ermöglicht es, einen Teil der Vorgänge zu standardisieren und damit zu automatisieren. Das schlägt sich in deutlich reduzierten Entwicklungszeiten nieder, die wiederum zu einem höheren Return on Investment führen.

Der Schichtaufbau der Data Vault-Modellierung



Die Data Vault-Architektur besteht im Wesentlichen aus drei Schichten:

Staging Layer	Data Warehouse Layer, bestehend aus				Information Mart Layer
Sie führt die Rohdaten aus den verschiedenen Datenquellen zusammen, etwa aus dem CRM- und ERP-System.	Raw Data Vault Hier sind die Rohdaten gespeichert.	Business Data Vault (optional) Die Daten sind auf der Basis von Geschäftsregeln harmonisiert.	Metrics Vault (optional) Hier befinden sich Laufzeitinformationen.	Operational Vault (optional) Hier fließen die Daten direkt aus operativen Systemen ein.	Hier findet die Modellierung statt und die Informationen werden für Analyse und Reporting einschließlich Visualisierung bereitgestellt.

### 3.3 Vorteile und Herausforderungen von Data Vault

#### Vorteile

- Es lassen sich Daten aus verschiedensten Quellsystemen miteinander integrieren. Zudem ist es einfacher als mit traditionellen Modellierungsmethoden, neue Datenquellen anzubinden.
- Die Daten lassen sich parallel, also voneinander unabhängig und zeitgleich, in den Speicher laden.
- Man kann das DWH dank seiner agilen Architektur sowohl sehr einfach skalieren als auch flexibel ausbauen und erweitern.
- Alle im DWH gespeicherten Informationen sind bis zur Datenquelle nachverfolgbar.
- Da die Rohdaten dauerhaft abgespeichert sind, ist Time Traveling möglich, das stichtagsbezogene Auswerten historischer Daten.
- Die ETL-Muster, nach denen Hubs, Links, und Satelliten geladen werden, sind einfach und einheitlich.

#### Herausforderungen

- Das zugrunde liegende Konzept ist komplex. Man benötigt deshalb spezielles Fachwissen, das ein Unternehmen in der Regel erst aufbauen muss.
- Ohne ein tiefergehendes Verständnis der geschäftlichen Zusammenhänge riskiert man, nur die Quelldaten zu kopieren und zu historisieren.
- Die Komplexität kann über Gebühr steigen, wenn die Anzahl der ETL-Prozesse sehr stark zunimmt. Der Grund: Durch Modellerweiterungen wird sich auch die Anzahl von Hubs, Links und Satelliten erhöhen.
- Das Laden von Data Marts ist zumeist umfangreich und komplex.

*Data Vault ist aktuell die wahrscheinlich leistungsfähigste Methode für ein skalierbares und agiles Modellieren der Daten im Core Data Warehouse.*

## 3.4 Praxisbeispiel: Unternehmensübernahme

Damit zwei Unternehmen fusionieren können, sind ihre Daten miteinander zu integrieren. Das Data Alignment muss zügig und vollständig erfolgen und umfasst alle Geschäftsdaten beider Unternehmen. Dazu gehören sämtliche – in der Regel auch die historischen – Daten aus dem Vertrieb, dem Personalwesen und dem kaufmännischen Bereich. Es geht also meistens um große Datenbestände.

Außerdem liegen die Datenobjekte in verschiedenen Systemen vor, notwendig ist daher eine plattformübergreifende Integration. Das gelingt am besten mit einer Data Vault-Lösung, zumal man damit in der Lage ist, agil auf Änderungen zu reagieren, die während des Merge-Prozesses jederzeit auftreten können. Die Rohdaten werden unverändert, das heißt entkoppelt von den Informationen abgespeichert, die erst beim Ausführen der Geschäftsregeln entstehen, und in Hubs, Links und Satelliten aufgeteilt.

## 3.5 Entscheidung leichter gemacht: Ist Data Vault für Sie geeignet?

Data Vault ist keine Universallösung für jedes Unternehmen. Denn sie zu konfigurieren, zu implementieren und bei Bedarf anzupassen, erfordert viel Know-how und ist entsprechend ressourcenintensiv. Sie sollten die Methode daher nur einsetzen, wenn Ihr Unternehmen von den beträchtlichen Vorteilen auch wirklich profitieren kann.

Das ist nicht der Fall, wenn

- in Ihrem Unternehmen nur geringe Datenmengen anfallen oder Sie die Daten nur einmalig nutzen,
- Sie auf nur eine oder sehr wenige Datenquellen zugreifen,
- es keinen Bedarf an Trendanalysen und stichtagsbezogenen Auswertungen gibt, und
- es keine oder nicht genügend Analysten gibt, welche die Ergebnisse interpretieren.

Data Vault ist die richtige Methode für Sie, wenn eine der folgenden Bedingungen für Ihr Unternehmen zutrifft:

- Es sind große Datenvolumen in kurzer Zeit zu laden.
- Es sollen Daten aus vielen verschiedenen Quellsystemen bezogen werden.
- Ein strategisches Ziel besteht darin, Business Intelligence-Applikationen agil zu entwickeln.
- Es ist geplant, ein vorgelagertes Core DWH innerhalb einer existierenden Silo-Architektur aufzubauen.
- Es ist gewünscht bzw. vorteilhaft, umfangreiche Datenauswertungen bis hin zu Time Traveling vorzunehmen.




### **Tipp: Data Vault praktisch umgesetzt mit Pentaho**

Das Pentaho Data Vault Framework erleichtert und beschleunigt den Aufbau eines agilen Data Warehouse. Durch Parametrisierung lassen sich per Drag & Drop neue Datenquellen einfach anbinden. Ebenso unkompliziert ist das Zusammenführen historisierter Daten und das Laden in die Data Marts. Im Vergleich zur herkömmlichen Vorgehensweise ist Data Vault 2-5-mal schneller.

## Sie möchten sich noch eingehender mit Data Vault beschäftigen?



Dann empfehlen wir Ihnen unser Whitepaper „Mit Data Vault zu mehr Agilität im Data Warehouse – Architekturen. Frameworks. Praxis.“

Es steht Ihnen  [hier](#) kostenlos zum Download bereit.

## 4. Modernisieren durch pseudonymisierte Personendaten

### 4.1 Ausgangslage

Organisationen haben großes Interesse daran, personenbezogene Daten in ihre Datenanalysen einzubeziehen. Doch es gibt einige Hürden, allem voran die rechtlichen Vorgaben der EU-Datenschutzgrundverordnung (DSGVO) und weiterer nationaler Gesetze zum Datenschutz, die jedoch selbst innerhalb der Europäischen Union variieren.

### Gründe, warum Unternehmen personenbezogene Daten analysieren möchten



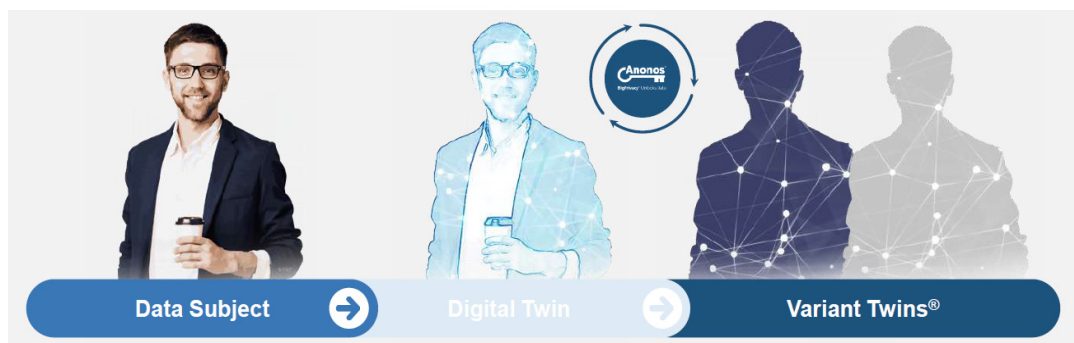
Unternehmen benötigen deshalb eine datenschutzkonforme Lösung, ohne dass ihre Data Warehouses an Flexibilität, Performance oder gar Nutzbarkeit verlieren.

#### Sie verarbeiten keine personenbezogenen Daten – sind Sie sicher?

Schließlich gehören nicht nur Namen, Adresse oder Geburtsdatum dazu. Als „personenbezogene Daten“ gelten alle Informationen zu einer Person, so auch Kontaktangaben, Bankverbindung oder Gesundheitsinformationen. Selbst Bestelldaten, IP-Adressen, Browser-Fingerprint, Verbindungsdaten und sogar Geodaten fallen darunter.

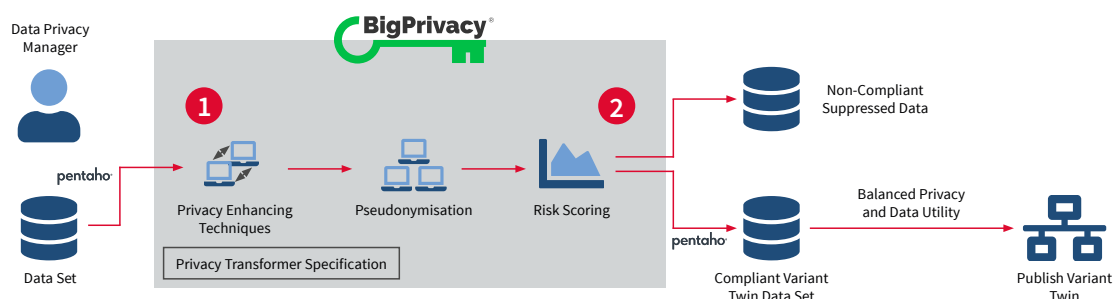
## 4.2 Lösung: Der abweichende Zwilling

Um unter Einhaltung der DSGVO personenbezogene Daten analysieren zu können, muss man die Daten beim Hochladen in das DWH pseudonymisieren. Dazu erstellt man einen Variant Twin, das bedeutet, man trennt die Informationen zum jeweiligen Thema von jenen Daten, die eine Einzelperson identifizieren.



*Pseudonymisierte Daten erlauben keine Rückschlüsse auf die natürliche Person. Zusätzliche Informationen wären dafür erforderlich, doch diese sind separat gespeichert und dürfen nur Personen mit Zugangsberechtigung zugänglich gemacht werden.*

Dynamische Entidentifizierer („Dynamic De-Identifiers“) zwischen und innerhalb der Datasets verhindern ein unautorisiertes Wiederverbinden der getrennten Informationen zu einer Person (Relinking). Die Rohdaten befinden sich im Data Privacy Server und sind für Anwender und Entwickler nicht zugänglich. Sie haben lediglich Zugriff auf die pseudonymisierten Daten im Data Mart.



*Mit dem Data Privacy Manager in Anonos kann man über sogenannte Privacy Actions einen Variant Twin erzeugen (1). Der K-Anonymity-Schwellenwert legt fest, welche Datensätze nicht in die Auswertung einfließen dürfen (2).*

### **Anonymisierung versus Pseudonymisierung**

Sind personenbezogene Daten komplett anonymisiert, lässt sich die dahinterstehende natürliche Person auf keinem Wege mehr identifizieren. Der Haken: Daten können bei der Anonymisierung verloren gehen und sind nie wieder herstellbar. Bei einer Pseudonymisierung dagegen können sie wieder hergestellt werden. Die Alternative ist die Pseudonymisierung. Hierzu ordnet man jeder Person ein Pseudonym zu. Dabei bleiben alle Information erhalten und dem Pseudonym zugeordnet. Die Nutzung von pseudonymisierten Daten erfüllt die Auflagen der DSGVO und ist ausdrücklich im Regelwerk vorgesehen.

## 4.3 Praxisbeispiel für pseudonymisierte Personendaten im Data Warehouse

Die medizinische und pharmazeutische Forschung hat großes Interesse daran, Behandlungs- und Patientendaten zu nutzen. Zum Beispiel werden solche Informationen für klinische Studien zur Arzneimittelzulassung genutzt, um Erkenntnisse zu Krankheitsbildern zu gewinnen oder um neue Behandlungsmethoden zu evaluieren. Dabei geht es nicht nur darum, dass ein Klinikum die Daten seiner eigenen Patienten auswertet, sondern sie für Forschungsk Kooperationen und in medizinischen Kompetenznetzen teilt.

Gesundheitsdaten sind jedoch besonders sensibel und dürfen deshalb nur so verwendet werden, dass sie keine Rückschlüsse auf die natürlichen Personen zulassen. Bei einer Anonymisierung ist das gewährleistet. Ebenso wird in der Praxis jedoch auch die Daten-Pseudonymisierung eingesetzt, etwa wenn die Datenerhebung in mehreren Etappen erfolgt oder sich über einen längeren Zeitraum erstreckt. Das ist zum Beispiel der Fall, wenn man Krankheitsverläufe betrachtet oder die langfristigen Auswirkungen eines Medikaments analysiert.

## 4.4 Ihre Entscheidung leichter gemacht: Verarbeitet Ihre IT die Daten DSGVO-konform?

Personenbezogene Daten, an deren Nutzung ein Unternehmen ein „legitimes Interesse“ hat, dürfen nach DSGVO weiterhin verarbeitet, analysiert und gespeichert werden – vorausgesetzt, sie sind pseudonymisiert und lassen keine Rückschlüsse auf persönliche Aspekte der einzelnen Person zu. Ist das in Ihrem Unternehmen der Fall?

Pseudonomysierte Daten liegen dann vor, wenn sich personenbezogene Daten nach ihrer Verarbeitung ohne zusätzliche Informationen keiner spezifischen Person mehr zuordnen lassen. Technische und organisatorische Maßnahmen müssen den Informationswert der Daten von den Mitteln zur Identifizierung der betroffenen Person trennen. Diese zusätzlichen Informationen sind deshalb auch getrennt zu speichern und dürfen nur Personen mit entsprechender Zugriffsberechtigung zugänglich sein.

## 5. So machen Sie Ihr Data Warehouse zukunftsfest

Traditionelle Data Warehouses sind nicht dafür ausgelegt, das explosive Datenwachstum von heute zu bewältigen, umfangreiche Analysen durchzuführen oder schnell und kostengünstig zu skalieren. Cloud-basierende DWH erfüllen aktuelle Analyseanforderungen und skalieren bei zunehmender Datenmenge automatisch.

Bezüglich Methodik, Architektur und Modellierung hat sich Data Vault bei der Modernisierung von DWH etabliert. Im Vergleich zu klassischen Ansätzen (3NF, Star Schema) wird die Agilität dramatisch verbessert: das DWH kann sukzessive erweitert werden und ist konzeptionell vorbereitet für Cloud und Real-Time.

Die Anonymisierung spiegelt einen veralteten Ansatz zum Datenschutz wider, der entwickelt wurde, als die Verarbeitung von Daten auf isolierte Anwendungen beschränkt war, bevor die Verarbeitung von „Big Data“ populär wurde. Die Pseudonymisierung, die sich derzeit auf dem neuesten Stand der Technik befindet, ermöglicht dagegen eine datenschutzgerechte Nutzung von Daten in der heutigen „Big Data“-Welt des Datenaustauschs und der Datenkombination.

### **So helfen wir Ihnen bei der DWH Modernisierung**

- Workshops zur Erarbeitung einer DWH-Modernisierungs-Strategie
- Durchführung von Architektur-Reviews und Unterstützung bei der Entwicklung von modernen Architekturen
- Beratung und Implementierung zum Thema Pseudonymisierung bei allen DWH Themen
- Proof of Concept / Proof of Value in allen Bereichen des DWH
- Beratung bei der Softwareauswahl und Lizenzierung
- Unterstützung und/oder Leitung von DWH Projekten



## Warum Sie mit it-novum sprechen sollten...

Wir setzen diese Business Intelligence- und Big Data-Vorteile gewinnbringend für Unternehmen um:

- 360-Grad Blick auf Ihre Kunden
- Fachabteilungen werten dank Self-Service Analytics Big Data-Daten selbst aus
- Identifikation neuer Umsatzquellen durch intelligente Nutzung von Unternehmensdaten
- Kosteneinsparung durch Einsatz eines Data Warehouse
- Vermeidung des aufwändigen und fehleranfälligen Excel-Chaos

Wenn Sie diese Vorteile auch in Ihrem Unternehmen nutzen wollen, sollten wir uns kennenlernen!

it-novum bietet umfangreiche Dienstleistungen für Ihr Big Data Analytics Projekt:

- Data Engineering
- Implementierung von Data Warehouses und Data Lakes
- Pentaho/SAP Connector für die Verarbeitung von SAP-Daten
- Pentaho/HVA Connector für die Analyse von Video Streams
- Predictive Analytics und Machine Learning
- Dashboards und Data Visualization
- Embedded Analytics

## Führend in Business Open Source-Lösungen und -Beratung

it-novum ist das führende IT-Beratungsunternehmen im deutschsprachigen Raum für Geschäftslösungen auf Basis von Open Source. Von unserem Hauptsitz in Fulda und den Niederlassungen in Düsseldorf, Dortmund, Wien und Zürich aus betreuen wir Mittelstandskunden und Großunternehmen sowie den öffentlichen Sektor.

Als Hitachi Vantara Preferred Partner für Big Data Insights und IoT sind wir Experten für den Einsatz von Pentaho. Mit unserer Expertise in Beratung, Training und Support helfen wir Unternehmen, aus ihren Daten Erkenntnisse zu gewinnen und ihre datengetriebenen Projekte erfolgreich zu gestalten.



### Ihr Ansprechpartner:

#### **Stefan Müller**

Director Big Data Analytics & IoT

✉ [stefan.mueller@it-novum.com](mailto:stefan.mueller@it-novum.com)

☎ +49 661 103 942

## Zu mehr Agilität und Flexibilität im Data Warehouse



**it-novum GmbH Deutschland**  
Hauptsitz Fulda: Edelzeller Straße 44 • 36043 Fulda  
Telefon: +49 661 103 333  
Niederlassungen in Düsseldorf & Dortmund

**it-novum Zweigniederlassung Österreich**  
Ausstellungsstraße 50 / Zugang C • 1020 Wien  
Telefon: +43 1 205 774 1041

**it-novum Schweiz GmbH**  
Hotelstrasse 1 • 8058 Zürich  
Telefon: +41 44 567 62 07

**it-novum.com**