Title:

DWH | Was ist ein Data-Warehouse?

Description:

Data-Warehouse-Systeme: Mit diesen Software-Lösungen realisieren Sie Ihre individuelle Data-Warehouse-Architektur mit minimalem finanziellem Aufwand.

Teaser

Immer mehr Unternehmen setzten auf dispositive Daten, um strategische Entscheidungen durch Berichte, Prognosen und Modelle abzusichern. Die Datenbasis wird dabei in der Regel von einem Informationssystem bereitgestellt, das separiert von operativen Datenverarbeitungssystem als zentrales Datenlager fungiert. Man spricht von einem Data-Warehouse, kurz DWH. Das DWH ist ein modular aufgebautes System, das diverse Werkzeuge der Datenerfassung, -aufbereitung und -speicherung umfasst. Wir geben Ihnen einen Einblick in den Aufbau der idealtypischen DWH-Architektur, stellen kommerzielle Anbieter etablierter Data-Warehouse-systeme vor und zeigen kostenlose Open-Source-Alternativen auf.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Data-Warehouse-Systeme: Mit Big-Data-Analysen zum Erfolg

Entscheidungsunterstützende Systeme haben in der Geschäftswelt eine lange Tradition. Bereits seit den 60er-Jahren nutzen Unternehmen Analysemethoden, um **dispositive Daten** zu gewinnen. Ziel ist es, dem Management mit datengestützten Berichten, Modellen und Prognosen bei der strategischen Ausrichtung der Geschäftsprozesse unter die Arme zu greifen.

\_TABLE\_OF\_CONTENT\_

Analytische Informationssysteme, die entsprechende Funktionalitäten zur Verfügung stellen, umfassen Konzepte wie **MIS** (Management Information Systems), **DSS** (Decision Support Systems) oder **EIS** (Executive Information Systems), die sich nur schwer voneinander abgrenzen lassen und seit den 90er-Jahren in der betrieblichen Praxis sowie bei der Vermarktung entsprechender Produkte unter dem Sammelbegriff **Business Intelligence (BI)** zusammengefasst werden.

[Hinweis: Bei Business Intelligence (BI) handelt es sich um einen Sammelbegriff für die IT-gestützte Aufbereitung und Analyse von Unternehmensrohdaten. BI soll Wissen generieren, das Entscheidungsträgern als Grundlage für die strategische Ausrichtung des Unternehmens dient.]

Die Datenbasis entscheidungsunterstützender Systeme im Rahmen der BI wird heute in der Regel von einem zentralen Datenlager, dem sogenannten **Data-Warehouse**, zur Verfügung gestellt. Wir führen Sie in die Grundlagen des Data-Warehousings ein, zeichnen die Referenzarchitektur eines solchen Informationssystems nach und stellen etablierte Anbieter kommerzieller DWH-Lösungen ebenso vor wie kostenlose Open-Source-Alternativen.

## Was ist ein Data-Warehouse?

Ein Data-Warehouse (DWH) ist gemäß Definition ein von operativen Datenverarbeitungssystemen separiertes Datenbanksystem, in dem Daten verschiedener, mitunter sehr heterogener Quellen zusammengeführt, verdichtet und langfristig archiviert werden. Zahlreiche Unternehmen überführen historische Daten operativer Datenverarbeitungssysteme in regelmäßigen Abständen in ein solches Datenlager und bereiten diese dort für spätere Zugriffe und strategische Analysen im Rahmen der **Business Intelligence (BI)** auf. Aus operativen Daten werden dispositive Daten:

* **Operative Daten:** Bei operativen Daten handelt es sich um transaktionsorientierte Informationen, die in Unternehmen während des Tagesgeschäfts anfallen und von Administrations- und Abrechnungssystemen generiert werden. Zu den typischen Datenquellen zählen operative Datenverarbeitungssysteme wie Buchhaltungsprogramme, Warenwirtschaftssysteme, Enterprise-Resource-Planning (ERP) oder Auskunfts- und Bestellsysteme.
* **Dispositive Daten:** Werden operative Daten an einer zentralen Stelle zusammengeführt (aggregiert), langfristig gespeichert und für Analysen aufbereitet, spricht man von dispositiven Daten.

Ein DWH bietet Analysten eine umfassende Sicht auf heterogene Datenbestände und ermöglicht die **Aggregation betrieblicher Kennzahlen** im Rahmen des **Online Analytical Processings (OLAP)**. Als zentrale Sammelstelle aller relevanten Unternehmensdaten dient das DWH dem unternehmensinternen **Wissensmanagement**. Anwendern wird dabei in der der Regel nur der Lesezugriff gewährt. Ein DWH fungiert als Datenbasis für **Data-Mining-Methoden** und ist Grundlage aller Überlegungen im Rahmen des Leistungsmanagements und der strategischen Unternehmensausrichtung.

## Aufbau eines DWH: Data-Warehouse-Architektur

Der Prozess der Bewirtschaftung und Auswertung eines DWH wird **Data-Warehousing** genannt und umfasst folgende Phasen:

Die Phasen des **Data-Warehousing** spiegeln sich im idealtypischen Aufbau, der sogenannten Referenzarchitektur, von Data-Warehouse-Systemen wider. Zwar unterscheidet sich die Systemarchitektur eines DWH je nach Produkt und Anbieter – grundsätzlich orientiert sich der technische Aufbau jedoch an einem modularen Muster, das sich in drei Ebenen gliedern lässt:

* **Datenerfassungsebene**
* **Datenhaltungsebene**
* **Datenbereitstellungsebene**

Zudem gibt es eine **zentrale Kontrollkomponente**: den Data-Warehouse-Manager, der jeder Ebene des DWH spezielle Administrationsfunktionen zuordnet. Die einzelnen Komponenten eines Data-Warehouse müssen nicht zwangsläufig von einem Anbieter stammen. Die jeweiligen Services können durchaus von unterschiedlichen Softwareprodukten oder Individuallösungen bereitgestellt werden.

Folgende Abbildung zeigt eine schematische Darstellung der DWH-Referenzarchitektur.

[Alt-Text: Die Data-Warehouse-Referenzarchitektur]

[Bildunterschrift: Die drei Ebenen der Data-Warehouse-Referenzarchitektur: Datenerfassung, Datenhaltung, Datenbereitstellung]

## Datenhaltungsebene

Bevor Daten ins DWH geladen werden können, müssen die oft sehr heterogenen Informationen in eine einheitliche Darstellungsform überführt werden. Ein DWH speist sich sowohl aus internen Datenquellen eines Unternehmens als auch aus relevanten externen Datenquellen:

* **Interne Daten:** Operative Systeme: Enterprise-Resource-Planning-Systeme (ERP), Customer-Relationship-Management-Systeme (CRM); operative Datenbanken; Content-Management-Systeme (CMS); Flat-Files (z. B. Excel, CSV, Text-Dateien), Mails etc.
* **Externe Daten:** Anwendungen und Systeme externer Dienstleister, Websites/Internet, Social Media, Cloud-Services etc.

Systeme auf der Datenerfassungsebene stellen Schnittstellen zu den operativen Systemen eines Unternehmens bereit und kommen in der ersten Phase des Data-Warehousings zum Einsatz: Zentrale Funktionen dieser DWH-Komponente sind **Datenbeschaffung** und **Datenintegration**

Im Rahmen der **Datenbeschaffung** kommen folgende Extraktionstechniken zum Einsatz:

* **Trigger:** Sofern die operativen Systeme eines Unternehmens Datenbanktrigger unterstützen, lassen sich diese einsetzen, um die Datenextraktion zu automatisieren. Trigger ermöglichen es, Operationen zu definieren, die beim Eintreffen bestimmter Ereignisse automatisch ausgeführt werden. In der Regel handelt es sich bei den auslösenden Ereignissen um Änderungen im Datenbestand der Quellensysteme, die zu einer Extraktion der modifizierten Daten in das DWH führen.
* **Protokolldateien:** Unterstützt ein operatives System die Trigger-Technologie nicht, kann die Datenerfassungsebene eines DWH Programme beinhalten, die in der Lage sind, die Protokolldateien (Logs) der Quellsysteme auszuwerten und die darin aufgezeichneten Operationen zu extrahieren.
* **Monitorprogramme:** Stehen für die Extraktion weder Trigger noch Protokolldaten zur Verfügung, kommen in der Regel Monitorprogramme zum Einsatz. Diese extrahieren Änderungen im Datenbestand eines operativen Systems anhand von Algorithmen, die in regelmäßigen Abständen Abbildungen (Snapshots) der zu überwachenden Daten erstellen und mit vorherigen abgleichen.

Wird keine der beschriebenen Techniken unterstützt, da der Zugriff auf den Datenbestand des operativen Systems nicht möglich ist, muss dieses Änderungen selbständig protokollieren und relevante Modifikationen an das Data-Warehouse übermitteln.

Die meisten DWHs stellen im Rahmen der Datenintegration **OLAP-Funktionalitäten** zur Verfügung, die es ermöglichen, Daten in mehrdimensionalen Strukturen darzustellen. Online Analytical Processing (OLAP) ist eine Analysemethode, die der Verdichtung managementrelevanter Unternehmensdaten dient. Das Verfahren beruht auf dem **ETL-Prozess**:

* **E = Extraction:** Die Datenextraktion umfasst das Auslesen relevanter Informationen aus verschiedenen Datenquellen. Das kann als Push- oder Pull- Strategie realisiert werden. Erfolgt die Datenextraktion im Rahmen einer Push-Strategie, werden Datenquellen dazu angehalten, in regelmäßigen Abständen Extrakte zu erzeugen und diese an das DWH zu übermitteln. Im Fall einer Pull-Strategie stößt das DWH die Datenextraktion von sich aus an.
* **T = Transformation:** Die extrahierten Daten werden im Rahmen einer Transformation bereinigt und einheitlich in das Format der Zieldatenbank übersetzt.
* **L = Loading:** Die Ladephase umfasst das Abspeichern der transformierten Daten in die jeweiligen Zieldatenbanken des DWH.

Die Datenerfassungsebene eines DWH kann eine sogenannte **Staging Area** beinhalten (auch Arbeitsbereich genannt). Dabei handelt es sich um einen temporären Bereich der Datenbank, in dem die Vorverarbeitung der zu ladenden Daten stattfindet. Ein solches Staging kann vor allem bei komplexen ETL-Prozessen erforderlich sein.

Da im DWH Daten unterschiedlichster Quellen zusammengeführt werden, stützt sich die **Datenintegration** auf diverse Werkzeuge, die eine Transformation und Bereinigung der extrahierten Daten ermöglichen. Diese lassen sich folgenden Kategorien zuordnen.

* **Data-Migration-Werkzeuge:** Programme für die Datenmigration ermöglichen es, einfache Transformationsregeln zu definieren, um heterogene Ausgangsdaten in ein einheitliches Zielformat zu überführen.
* **Data-Scrubbing-Werkzeuge:** Im Rahmen des Data-Scrubbings kommen Programme zum Einsatz, die sich auf den Fuzzy-Logic-Ansatz und Neuronale Netze stützen. Ziel ist die Verbesserung der Datenqualität, indem Fehler, Lücken und Wiederholungen in Datensätzen durch vordefinierte Regeln, Algorithmen oder Lookup-Tables (LUT) ausgebessert werden. Man spricht in diesem Fall auch von *Quality Management*.
* **Data-Auditing-Werkzeuge:** Data-Auditing-Werkzeuge kommen im Rahmen der Datenintegration zum Einsatz, um Regeln und Beziehungen zwischen Daten zu ermitteln. Zudem ermöglichen Programme dieser Art, Daten zu identifizieren, die gegen die ermittelten Regeln verstoßen und somit voraussichtlich fehlerhaft sind.

Auf die Datenintegration folgt die Übernahme der extrahierten Daten in die zentrale Datenbank, das sogenannte **Core-Data-Warehouse**. Unterstützt wird dieser Schritt durch Programme, die folgende Funktionen zur Verfügung stellen:

* Integritätsbedingungen prüfen
* Daten sortieren
* Aggregationen berechnen
* Zugriffsstrukturen berechnen
* Daten für einen effizienten Zugriff partitionieren

### Datenhaltungsebene

Kern des DWH stellt die Datenhaltungsebene dar. Diese umfasst das sogenannte Core-Data-Warehouse. Extrahierte Daten werden im DWH meist in Form von mehrdimensionalen Matrizen, sogenannten **Stern- und Schneeflockenschemata,** gespeichert und im Rahmen der Langzeitarchivierung dauerhaft für zukünftige Analysen vorgehalten. Diese beziehen sich jedoch nur selten auf den gesamten Datenbestand des DWH. Um eine effiziente Auswertung zu ermöglichen, ist es daher üblich, Datenausschnitte des Gesamtbestands, sogenannte **Data-Marts** anzulegen.

Ein Data-Mart ist eine Kopie eines Teildatenbestandes, die in der Regel nicht-persistent als Zwischenspeicher realisiert wird. Mitunter werden jedoch auch sogenannte **unabhängige Data-Marts** genutzt, die einen separaten Datenausschnitt dauerhaft vorhalten.

Bei einem **Sternschema** handelt es sich um eine Form eines **Entitäten-Relationen-Diagramms** (*Entity Relationship Diagram*, ERD) – also eine grafische Darstellung der Tabellenstruktur einer Datenbank, bei der die verschiedenen Entitäten sowie deren Beziehungen zueinander veranschaulicht werden. Das Sternschema dient somit der Visualisierung multidimensionaler Datenstrukturen.

Jedes Sternschema besteht aus einer Faktentabelle und mehreren Dimensionstabellen, die sich eben **sternförmig** um die Faktentabelle gruppieren.

* Die **Faktentabelle** beinhaltet sogenannte Fakten: Kenn- und Ergebniszahlen eines Unternehmens, die fortlaufend festgehalten werden (beispielsweise der Umsatz).
* Die **Dimensionstabellen** enthalten Attribute, mit denen sich die Daten der Faktentabelle beschreiben lassen. Es handelt sich bei einer Dimensionstabelle somit um eine Sammlung von Referenzinformationen zu den in der Faktentabelle gespeicherten Ereignissen.

In einem Sternschema (*star schema*) ist lediglich die Faktentabelle mit allen Dimensionstabellen über Fremdschlüsselbeziehungen verbunden. Verbindungen zwischen einzelnen Dimensionstabellen werden nicht hergestellt. Folgende Abbildung zeigt eine vereinfachte Darstellung einer solchen Datenstruktur:

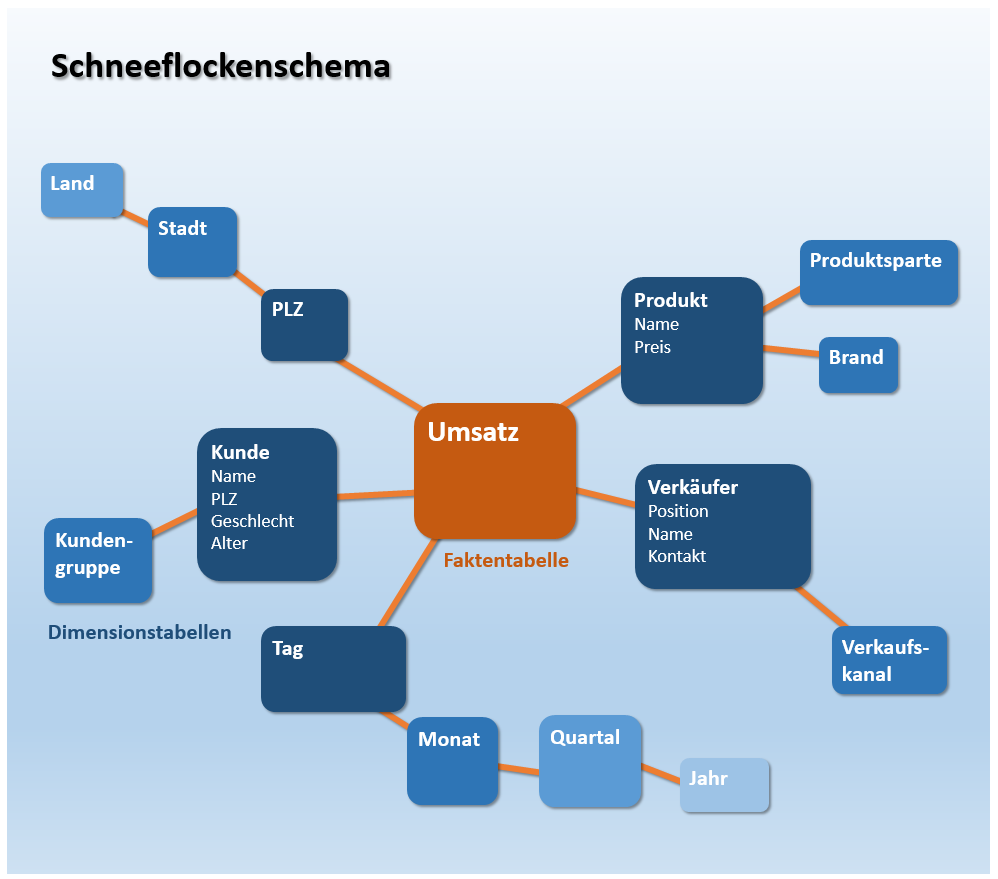


[Alt-Text: Beispiel eines Sternschemas]

[Bildunterschrift: Um die Faktentabelle *Umsatz* gruppieren sich die Dimensionstabellen *Produktsparte*, *Produkt*, *Verkaufskanal*, *Zeitraum*, *Kunde* etc.]

Im oben dargestellten Sternschema lässt sich der Fakt *Umsatz* beispielsweise in Relation zu einem bestimmten *Verkaufskanal*, einem speziellen *Produkt*, einem *Verkäufer*, der *Region* oder dem jeweiligen *Zeitraum* darstellen. Eine mögliche Abfrage über drei Dimensionen wäre beispielsweise: Wieviel Umsatz wurde mit einem bestimmten Produkt im Zeitraum 2016 über den Verkaufskanal *Online* erzielt?

Eine Weiterführung des Sternschemas ist das **Schneeflockenschema** (*snowflake schema*). Während die Dimensionstabellen eines Sternschemas in denormalisierter Form vorliegen, werden Referenzinformationen beim Schneeflockenschema gemäß der 3. Normalform gespeichert. Es findet somit eine Klassifizierung und Hierarchisierung der Daten statt, bei der redundante Informationen in neue Tabellen ausgelagert werden. Es entstehen dadurch die charakteristischen Verzweigungen, die an Schneeflocken erinnern.



[Alt-Text: Beispiel eines Schneeflockenschemas]

[Bildunterschriften: Im Schneeflockenschema werden redundante Informationen in separate Dimensionstabellen ausgelagert, klassifiziert und hierarchisiert]

Schneeflockenschemata zeichnen sich gegenüber Sternschemata durch einen geringeren Speicherplatzverbrauch aus. Dieser resultiert aus der normalisierten Datenhaltung. **Unter Normalisierung versteht man die Auslagerung von Tabellenspalten in neue Tabellen mit dem Ziel, doppelte Einträge zu vermeiden.** Der Abbau von Redundanzen reduziert zudem den Aufwand im Rahmen der Datenpflege: Jede Information liegt im besten Fall nur einmal vor und muss somit bei Bedarf nur an einer Stelle angepasst werden.

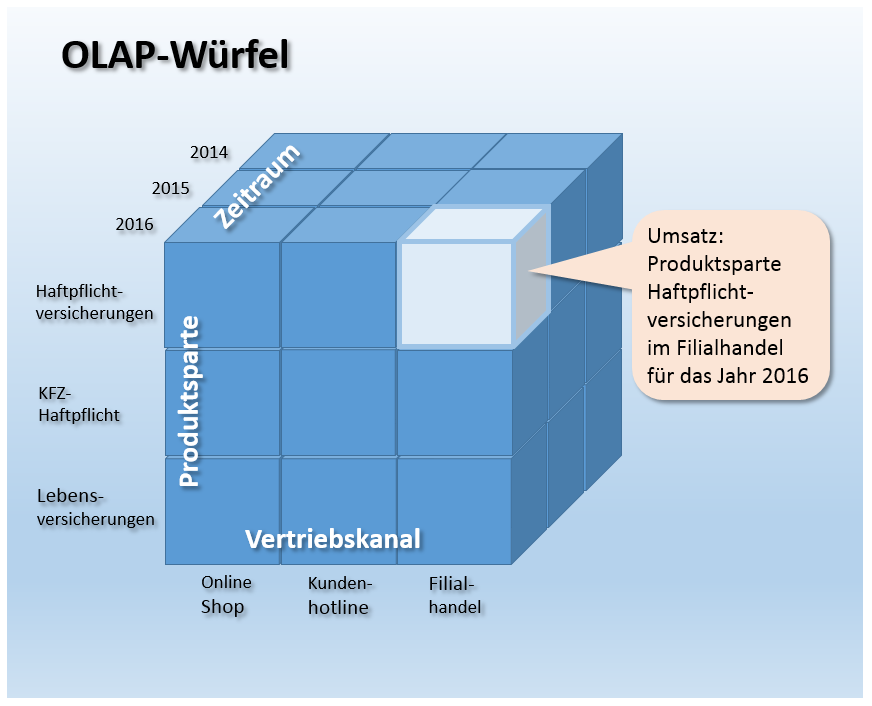
Werden Daten in normalisierte Tabellen ausgelagert, führt dies jedoch zwangsläufig zu komplexeren Datenstrukturen, die in der Regel längere Abfragezeiten mit sich bringen. Möchten Analysten auf Daten in einem Schneeflockenschema zugreifen, müssen die mehrstufigen Dimensionstabellen zunächst im Rahmen eines **Joins** (Verbund) verknüpft werden.

[Hinweis: bei einem Join handelt es sich um eine Datenbankoperation, mit der sich via Fremdschlüssel verknüpfte Datenbanktabellen unter bestimmten Bedingungen zusammenführen lassen.]

In der Praxis basiert die Datenstruktur eines DWH meist auf dem Schneeflockenschema, während einzelne Data-Marts als Sternschemata realisiert werden.

Man spricht bei Stern- bzw. Schneeflockenschemata von Dimensionstabellen, da sich jede Tabelle als eine Dimension eines mehrdimensionalen **OLAP-Würfels** darstellen lässt. Dies ermöglicht es Analysten, die im DWH gespeicherten Fakten in Relation zu beliebig vielen Referenzinformationen zu setzen, um betriebswirtschaftliche Kennzahlen – wie beispielsweise den Umsatz – anhand verschiedener Aspekte mehrdimensional zu analysieren und in diversen Detaillierungsstufen zu untersuchen.

Folgende Abbildung zeigt die schematische Darstellung eines dreidimensionalen OLAP-Würfels, dessen Kanten die Dimensionen *Produktsparte*, *Vertriebskanal* und *Zeitraum* aufspannen. Die Länge der Würfelkanten wird durch die Anzahl der Zellen bestimmt. Jede Würfelzelle enthält genau einen Wert – beispielsweise den Umsatz für die Produktsparte *Haftpflichtversicherungen* im Jahr *2016* über den Vertriebsweg *Filialhandel* (in der Grafik hellblau hervorgehoben).



[Alt-Text: Schematische Darstellung eines OLAP-Würfels]

[Bildunterschrift: OLAP-Würfel mit drei Dimensionen]

Das OLAP-Verfahren ist nicht auf drei Dimensionen beschränkt. Ein solcher Datenwürfel ist **n-dimensional** aufgebaut und kann prinzipiell beliebig viele Dimensionen umfassen.

[Hinweis: Je nachdem, welche Speichertechnik dem Core-Data-Warehouse zugrunde liegt, unterschiedet man zwischen verschiedenen OLAP-Verfahren. Greift der Würfel auf Daten aus einer relationalen Datenbank zurück, spricht man von ROLAP (relationales OLAP). Würfel auf Basis multidimensionaler Datenbanken werden MOLAP (multidimensionales OLAP) genannt.]

### Datenbereitstellungsebene

Die Datenbereitstellungsebene fungiert als Schnittstelle zu Endanwendungen und Präsentationswerkzeugen. Methoden zur **Datenauswertung und -analyse** werden von diversen Endanwenderwerkzeugen zur Verfügung gestellt. Diese ermöglichen es, Informationen aus dem Datenbestand des Data-Warehouse zu extrahieren und für Endanwender in unterschiedlichen Darstellungsformen aufzubereiten. Das Spektrum umfasst Bericht- und Abfragewerkzeuge, Kollaborations-Tools, Data-Mining-Werkzeuge, Werkzeuge des Online Analytical Processing (OLAP), Executive Information Systems (EIS) sowie Forecasting- und Simulation-Tools.

#### Bericht- und Abfragewerkzeuge

Berichtwerkzeuge stellen Endanwendern unterschiedliche Funktionen zur Verfügung, um **vordefinierte Standardberichte** (*Predefined Reportings*) zu erstellen. Dies kann automatisiert in regelmäßigen Zeitabständen erfolgen oder bei Bedarf auf Anfrage. Um Endanwendern Anfragen an das DWH zu erleichtern, lassen sich auch diese mithilfe von Anfragewerkzeugen vordefinieren.

#### Kollaborations-Tools

Kollaboration-Tools unterstützen die Kommunikation und **Zusammenarbeit von Endanwendern** bei der Datenanalyse. Das Funktionsspektrum dieser Werkzeuge umfasst beispielsweise das Speichern von Anmerkungen und den Austausch von Analyseergebnissen.

#### Data-Mining-Werkzeuge

Unter den Sammelbegriff **Data-Mining** [Data-Mining: Analysemethoden für Big Data] (https://hosting.1und1.de/digitalguide/online-marketing/web-analyse/data-mining-so-lassen-sich-datenberge-erschliessen/) fallen alle ungerichteten, teilweise automatisierten Analysemethoden, die darauf abzielen, relevante Muster, Trends und Beziehungen im Datenbestand zu ermitteln. Data-Mining-Werkzeuge stützen sich auf statistische und mathematische Methoden sowie auf Techniken der Künstlichen Intelligenz (KI) und des Maschinenlernens. Der Umfang der Daten, die Unternehmen erzeugen, verarbeiten und zu Analysezwecken in DWHs zusammenführen, wächst exponentiell. Das durchschnittliche Datenvolumen weltweit verdoppelt sich alle zwei Jahre. Vor diesem Hintergrund gewinnen Data-Mining-Methoden im Rahmen des Data-Warehousings zunehmend an Bedeutung.

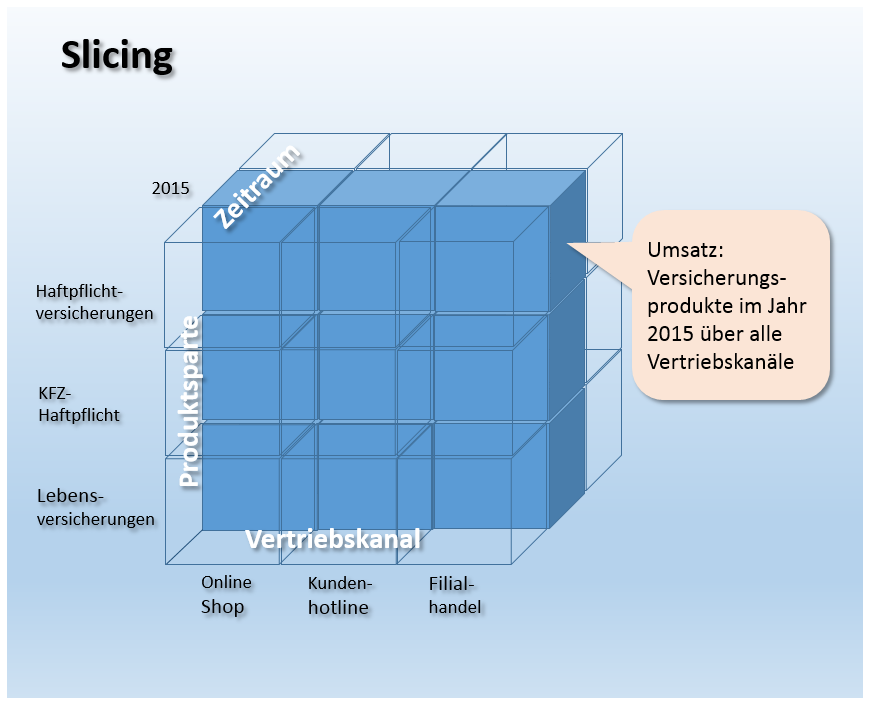
#### Werkzeuge des Online Analytical Processing (OLAP)

Von den zur Verfügung stehenden Datenauswertungs- und Analyse-Werkzeugen haben sich im Rahmen des Data-Warehousing vor allem **OLAP-Anwendungen als Standardbenutzerschnittstelle** etabliert. Werkzeuge, die im Rahmen des Online Analytical Processings zum Einsatz kommen, stellen Endanwendern verschiedene Funktionen zur Verfügung, mit denen sich Anfragen an das DWH ad-hoc formulieren lassen. Sie dienen der Navigation durch den multidimensionalen Datenbestand.

Die Darstellung via OLAP ermöglicht es, aufbereitete Daten in Abhängigkeit zu beliebig vielen vordefinierten Dimensionen zu modellieren. Analysten stehen dabei verschiedene **Grundoperationen** zur Verfügung, mit denen sich ein OLAP-Würfel bearbeiten lässt.

* **Slicing:** Als Slicing bezeichnet man ein Verfahren, das eine Dimension des OLAP-Würfels auf eine Teilmenge eingegrenzt. Es wird praktisch eine Scheibe aus dem Datenwürfel herausgeschnitten und separat betrachtet.

In folgender Abbildung wurde die Dimension *Zeitraum* auf die Teilmenge 2015 eingegrenzt. Der Ausschnitt zeigt somit die Umsätze für alle Versicherungsprodukte, die im Jahr 2015 über alle Vertriebskanale generiert wurden.

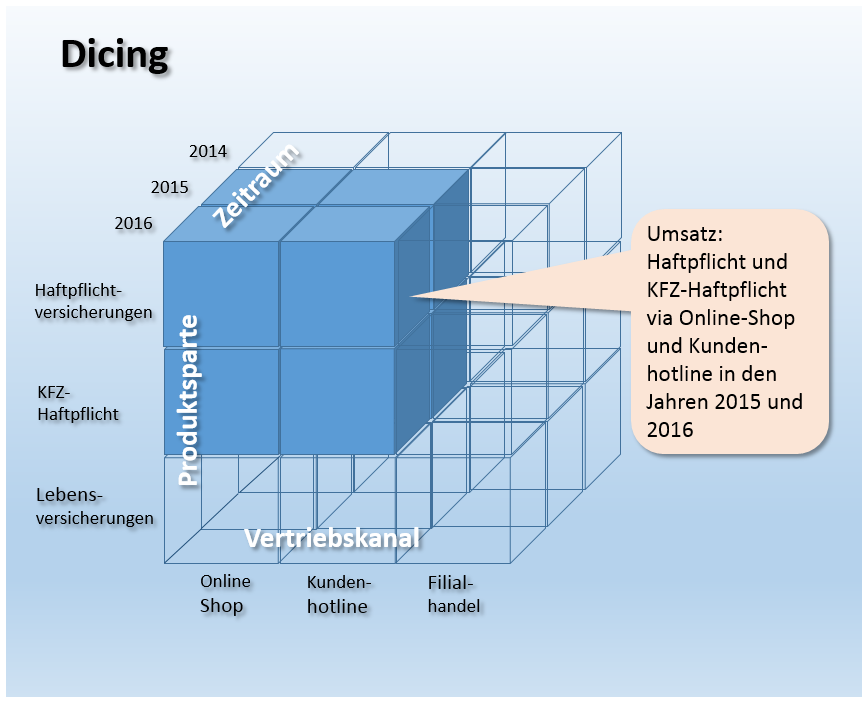


[Alt-Text: Schematische Darstellung einer Slicing-Operation am Beispiel eines dreidimensionalen OLAP-Würfels.]

[Bildunterschrift: Datenausschnitt eines dreidimensionalen OLAP-Würfels via Slicing]

* **Dicing:** Wird ein OLAP-Würfel durch simultane Slicing-Operationen in mehreren Dimensionen beschnitten, spricht man von Dicing. Beim Dicing-Verfahren wird ein kleinerer Würfel erzeugt, der eine Teilmenge des Gesamtwürfels darstellt.

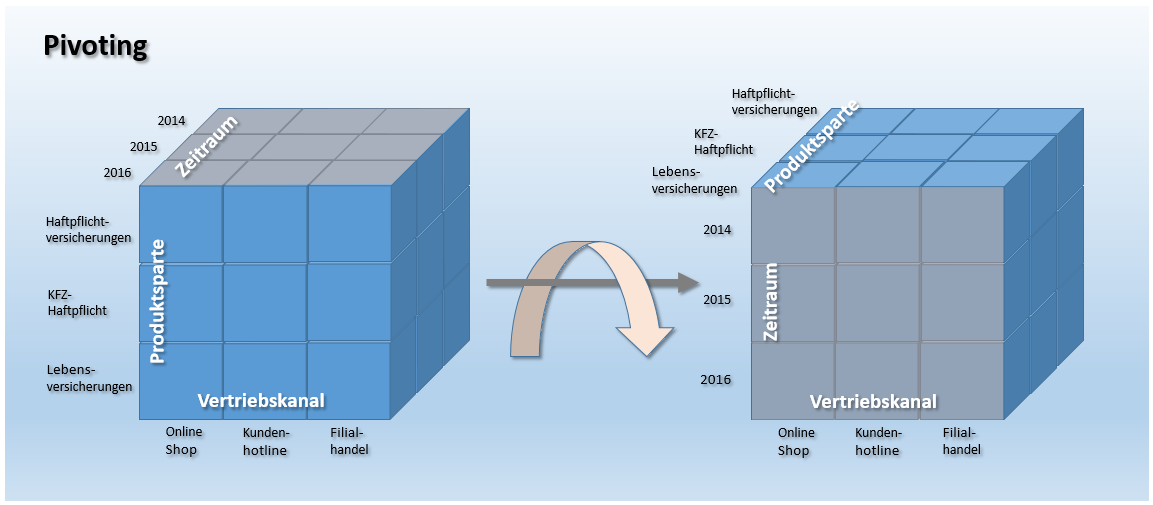
Folgende Abbildung zeigt eine Dicing-Operation, bei der der Gesamtwürfel in allen drei Dimensionen auf eine Teilmenge reduziert wurde.



[Alt-Text: Schematische Darstellung einer Dicing-Operation am Beispiel eines dreidimensionalen OLAP-Würfels]

[Bildunterschrift: Die modellierten Umsatz-Daten wurden in drei Dimensionen beschnitten]

* **Pivoting:** Das Drehen des Datenwürfels, sodass mindestens eine andere Dimension sichtbar wird, bezeichnet man als Pivoting.

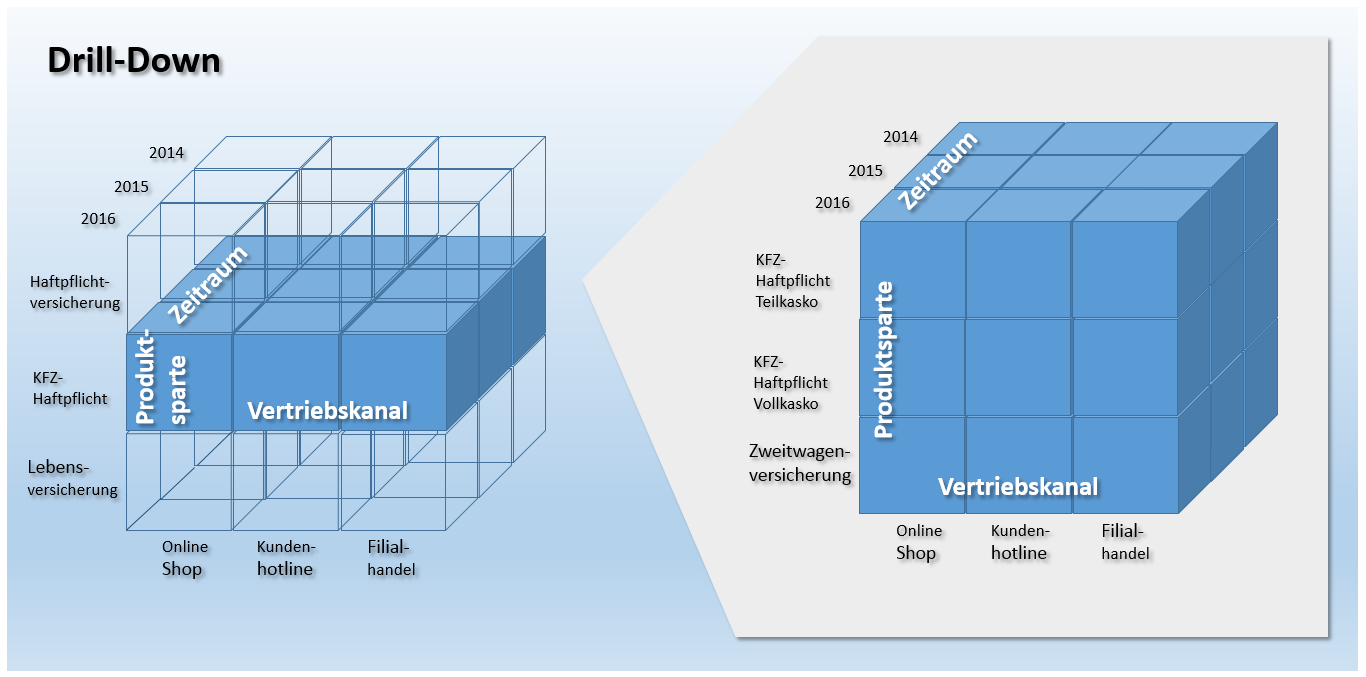


[Alt-Text: Schematische Darstellung einer Pivoting-Operation am Beispiel eines dreidimensionalen OLAP-Würfels]

[Bildunterschrift: Im Rahmen des Pivotings wird der OLAP-Würfel so rotiert, dass mindestens eine andere Dimension sichtbar wird]

* **Drill-Down/Roll-Up:** Sollen die Aggregationen eines Informationsobjekts auf detailliertere Werte heruntergebrochen werden, kommt die Operation Drill-Down zum Einsatz. Diese ermöglicht es Analysten, in einen OLAP-Würfel hineinzuzoomen und somit die Granularität der Daten zu erhöhen. Als Gegenoperation zum Drill-Down dient ein Roll-Up dem Verdichten von Informationen auf höhere Hierarchiestufen. Drill-Down und Roll-Up kommen bei der Navigation in mehrdimensionalen hierarchischen Strukturen zum Einsatz.

Folgende Abbildung zeigt ein Drill-Down des Informationsobjekts *Umsatz* in der Dimension *Produktsparte*. Die Granularität wird erhöht, sodass sich die im DWH gespeicherten Umsatzzahlen in Bezug auf einzelne Produkte interpretieren lassen.



[Alt-Text: Schematische Darstellung einer Drill-Down-Operation am Beispiel eines dreidimensionalen OLAP-Würfels]

[Bildunterschrift: Drill-Down der Dimension *Produktsparte* auf die Ebene einzelner Produkte]

* **Drill-Out/Split:** Der Drill-Out-Operator (auch Split genannt) ermöglicht es Analysten, einem OLAP-Würfel weitere Dimension hinzuzufügen. Das Ergebnis sind detailliertere Daten. Anders als beim Drill-Down wird der Detailgrad jedoch nicht in Bezug auf die Granularität erhöht, sondern durch einen Informationsgewinn, der aus den zusätzlichen Referenzinformationen der hinzugefügten Dimensionen resultiert.
* **Drill-In/Merge:** Als Gegenoperation zum Drill-Out wird der Detailgrad des OLAP-Würfels beim Drill-In durch das Entfernern von Dimensionen verringert. Im Gegensatz zu Roll-Up resultiert der Informationsverlust auch bei dieser Operation nicht aus einer Veränderung der Betrachtungsebene, sondern aus dem Verlust dimensionaler Informationen. Die Granularität bleibt die Gleiche.
* **Drill-Across:** Auch die Datenoperation Drill-Across dient der Analyse des Datenbestandes. Während sich die bisher besprochenen Operationen stets auf einen OLAP-Würfel beziehen, wird das Drill-Across-Verfahren auf mehrere korrelierte Datenwürfel angewendet, um globale Analysen zu ermöglichen. Dabei werden beliebig viele Faktentabellen mit mindestens einer gemeinsamen Dimension auf der gleichen Hierarchiestufe und Granularität (sprich: unter Beibehaltung der Betrachtungsebene) analysiert.
* **Drill-Through:** Bei einem Drill-Through handelt es sich um eine Operation, bei der ein Analyst eine einzelne Zelle eines Datenwürfels auswählt und **im höchsten Detaillierungsgrad** betrachtet. Anders als beim Drill-Down greift der Drill-Through dabei auf die Quelldaten der ausgewählten Würfelzelle zu. Das Resultat der Drill-Through-Operation wird somit aus den Tabellenzellen abgeleitet, die der Berechnung der ausgewählten Würfelzelle zugrunde liegen.

#### Executive Information Systems (EIS)

Ähnlich wie OLAP stellen EIS-Tools Endanwendern verschiedene Möglichkeiten zur Verfügung, Ad-hoc-Anfragen zu formulieren und Daten zu modellieren. In Abgrenzung zu OLAP wird der Begriff EIS heutzutage jedoch in erster Linie auf fertige Anwendungssysteme bezogen, die vordefinierte Berichte für bestimmte Betriebsbereiche wie den Verkauf, das Marketing oder die Finanzplanung zur Verfügung stellen.

#### Forecasting- und Simulation-Tools

Forecasting- und Simulation-Tools bieten Endanwendern die Möglichkeit, im DWH gespeicherte Kennzahlen in die Zukunft fortzuschreiben, um Vorhersagemodelle zu erstellen.

### Data-Warehouse-Management

Auf allen Ebenen des DWHs sind spezielle Werkzeuge aktiv, die im Bereich **Warehouse-Management** zusammengefasst werden. Aufgabe dieser Komponenten ist der Aufbau, die Wartung und der Betrieb aller Administrationsfunktionen, die im Rahmen des Data-Warehousings benötigt werden. Zentrale Aufgabenfelder des DWH-Managers sind das Scheduling der DWH-Prozesse, das Metadaten-Management, das Sicherheitsmanagement sowie das Systemmanagement.

* **Scheduling:** Das Scheduling umfasst die Steuerung der DWH-Prozesse. Administrationsfunktionen im Rahmen des Schedulings lassen sich in Bezug auf die Ebenen der Data-Warehouse-Architektur folgendermaßen kategorisieren:
  + **Datenerfassung/Datenintegration:** Auf der Datenerfassungsebene ist der DWH-Manager für das Design und die Anpassung der ETL-Prozesse zuständig. Darüber hinaus werden Administrationsfunktionen bereitgestellt, um Aktualisierungsvorgänge und das Qualitätsmanagement zu überwachen.
  + **Datenhaltung:** Auf der Ebene der Datenhaltung überwacht der DWH-Manager die Speicherauslastung, konstruiert Aggregationstabellen und führt Archivierungs- und Backup-Operationen aus.
  + **Datenbereitstellung:** Administrationsfunktionen auf der Datenbereitstellungsebene umfassen die Benutzerverwaltung sowie die Überwachung von Anfragelaufzeiten.
* **Metadaten-Management**: Zentrale Komponente des DWH-Managers ist das Metadaten-Repository. Es enthält alle Informationen, die für die Konstruktion und den Betrieb des DWHs erforderlich sind, sowie Informationen über den Datenbestand des DWH. Im Repository gespeicherte Metadaten umfassen beispielsweise die Definition des zugrundeliegenden Datenbank-Schemas, Informationen zu Speicherstrukturen, zu Zugriffspfaden und Dateigrößen, Metadaten zur Beschreibung der Datenquellen sowie Aktualisierungszeitpunkte, Datenbereinigungs- und Transformationsregeln, Indizes und Partitionstabellen. Darüber hinaus sorgt der DWH-Manager für einen Austausch der Metadaten zwischen den einzelnen Komponenten des DWH und stellt somit eine homogene Metadatenbasis bereit.
* **Sicherheitsmanagement:** Das Sicherheitsmanagement umfasst diverse Dienste im Rahmen der Nutzerauthentifizierung, Autorisierung und Verschlüsselung.
* **Systemmanagement:** Im Rahmen des Systemmanagements stellt der DWH-Manager verschiedene Administrationsfunktionen für den Betrieb des DWH bereit. Diese umfassen beispielsweise das Monitoring (Performance, Auslastung etc.), die Datenarchivierung oder die Datensicherung.

## Data-Warehousing unter datenschutzrechtlichen Gesichtspunkten

Die groß angelegte Aggregation von Betriebs-, Geschäfts- und Kundendaten in einem Data-Warehouse sowie die Analyse dieser Datenmassen mithilfe von Data-Mining-Methoden oder OLAP-Anwendungen bieten für Unternehmen Chancen, Geschäftsprozesse nachhaltig zu optimieren. Datenschützer betonen neben den Vorteilen im Rahmen der Entscheidungsfindung jedoch auch die Risiken solcher Big-Data-Analysen – speziell für das **Grundrecht auf informationelle Selbstbestimmung** und für den Schutz der Privatsphäre.

Heikel seien laut Kritikern vor allem Analysen, die die **Erstellung von Persönlichkeitsprofilen** und **automatisierte Vorhersagen von Verhaltens- und Handlungsweisen** ermöglichen. Im Fokus der Debatte steht das Manipulationspotenzial von Informationen, die aus Data-Analysen gewonnen wurden.

Eine offizielle Stellungnahme zum Themenfeld „Data-Warehouse, Data Mining und Datenschutz“ bietet die Entschließung der 59. Konferenz der Datenschutzbeauftragten des Bundes und der Länder. In dieser formulieren die Datenschützer folgende Rahmenbedingungen, die für eine rechtskonforme Speicherung personenbezogener Daten sowie deren Weiterverarbeitung gelten:

* **Zweckbindung personenbezogenen Daten:** Personenbezogene Daten dürfen nur im Rahmen der gesetzlich zugelassenen Zwecke oder der gegenseitigen Vereinbarungen erfasst, gespeichert und verarbeitet werden. Eine Speicherung personenbezogener Daten in einem Data-Warehouse entfernt sich den Datenschützern zufolge vom ursprünglichen Verwendungszweck und stellt eine unzulässige Speicherung auf Vorrat ohne Zweckbindung dar.
* **Zweckänderung nur mit Einwilligung:** Eine Änderung des Speicherzwecks personenbezogener Daten ist nur mit Einwilligung der Betroffenen zulässig. Diese müssen zudem über die Tragweite der Einwilligung aufgeklärt werden. Verfahren, bei denen personenbezogene Daten erhoben werden, müssen so gestaltet werden, dass Betroffene die Risiken abschätzen und ihre Rechte wahrnehmen können. Eine Einwilligung kann jederzeit zurückgezogen werden.
* **Erhebung personenbezogener Daten nur, wenn nötig:** Datenverarbeitungssysteme sind so zu gestalten, dass so wenig personenbezogene Daten wie nötig erhoben werden. Anonymisierte und pseudonymisierte Verfahren sind den Datenschützern zufolge unbedenklich.
* **Die permanente Speicherung personenbezogener Daten ist unzulässig:** Bei der Speicherung personenbezogener Daten sind die gesetzlichen Speicherfristen zu beachten. Eine Speicherung über die gesetzlichen Fristen hinaus ist rechtswidrig.
* **Automatisierte Einzelentscheidungen sind zu vermeiden:** Mitunter werden Data-Mining-Verfahren im Rahmen automatisierter Einzelentscheidungen verwendet. Dabei handelt es sich um Entscheidungen, die sich ausschließlich auf eine automatisierte Verarbeitung personenbezogener Daten stützen, die der Bewertung einzelner Persönlichkeitsmerkmale dient. Ein solches Vorgehen ist unzulässig: Die europäische Datenschutzrichtlinie spricht jeder Person das Recht zu, keiner belastenden automatisierten Einzelentscheidung unterworfen zu werden.

Die Entschließung der Datenschutzbeauftragten endet mit der Empfehlung an Hersteller und Anwender, bei Data-Warehouse-Systemen und Data-Mining-Verfahren grundsätzlich auf datenschutzfreundliche Technologien zu setzen, die eine **Speicherung personenbezogener Daten durch Anonymisierung oder Pseudonymisierung** vermeiden.

## Data-Warehouse-Software

Data-Warehousing ist schon lange kein Unterfangen mehr, das lediglich in den Chefetagen großer Konzerne diskutiert wird. Auch **kleine und mittlere Unternehmen (KMU)** sehen das Potenzial, durch Big-Data-Analysen Geschäftsprozesse zu optimieren. Neben hochpreisigen BI-Suiten und integrierten DWH-Komplettlösungen sind seit einigen Jahren daher auch kostengünstige Einsteigerprodukte, flexible Cloud-Services und ausgereifte Open-Source-Anwendungen auf dem Markt, die speziell den Mittelstand ins Visier nehmen.

### Kostenpflichtige Data-Warehousing-Produkte

Etablierte kommerzielle BI-Software zeichnet sich in der Regel durch eine hohe Zuverlässigkeit, ein im Rahmen von Service-Level-Agreements (SLA) vereinbartes Leistungsspektrum und einen professionellen Support aus. Anwender müssen dafür Kosten bei der Anschaffung oder für die cloudbasierte Nutzung einplanen.

Folgende Liste zeigt einen **Überblick kostenpflichtiger Data-Warehousing-Produkte** führender Anbieter in alphabetischer Reihenfolge.

|  |  |
| --- | --- |
| **Anbieter proprietärer Software** | **Data-Warehousing-Produkte** |
| **Amazon Web Services** | Amazon Redshift |
| **Cloudera** | Cloudera Enterprise |
| **Hewlett Packard Enterprise** | HP Vertica  HP ArcSight Data-Platform  HP Haven OnDemand  HP IDOL  HP Key View |
| **IBM** | IBM Netezza  IBM PureData System  IBM InfoSphere DataStage |
| **Microsoft** | SQL Server  Microsoft Analytics Platform System  Azure HDInsight for Hadoop |
| **Oracle** | Oracle Business Intelligence  Oracle Database  Oracle Exadata Database Machine  Oracle NoSQL Database  Oracle TimesTen In-Memory Database  Oracle Big Data Appliance |
| **Pivotal Software** | Pivotal Greenplum  Pivotal Big Data Suite  Pivotal HDB (powered by Apache HAWQ)  Pivotal HDP (OEM Hortonsworks Data Platform) |
| **SAP** | SAP NetWeaver Business Intelligence  SAP IQ  SAP HANA Enterprise Cloud |
| **SAS** | SAS Data Management  SAS Access Interface to Hadoop  SAS Federation Server  SAS Data Loader for Hadoop  SAS Event Stream Processing |
| **Snowflake Computing** | Snowflake |
| **Teradata** | Teradata Active Enterprise Data Warehouse  Teradata Data Warehouse Appliance  Teradata Appliance for Hadoop  Teradata Integrated Big Data Platform  Teradata Aster Big Analytics Appliance |

### Open-Source-Lösungen

Neben kostenpflichtigen Produkten bietet der Markt für Business-Intelligence-Software diverse Open-Source-Lösungen, die **Data-Warehousing-Funktionalitäten kostenlos** zur Verfügung stellen. Folgende Übersicht entstammt der Publikation „Business-Intelligence-Werkzeuge: Marktübersicht Open-Source-Werkzeuge aus dem Bereich Business Intelligence [Die BMWI-Publikation als PDF via mittelstand-digital.de] (http://www.mittelstand-digital.de/DE/wissenspool,did=680480.html)“, die das Bundesministerium für Wirtschaft und Energie (BMWI) über die Website *www.mittelstand-digital.de* zur Verfügung stellt. Die Tabelle zeigt die geläufigsten **BI-Open-Source-Programme** sowie deren Anwendungsgebiete.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **BI-Software** | **Rohdaten extrahieren** | **Rohdaten transformieren** | **Trans-formierte Daten laden** | **OLAP** | **Data Mining** | **Dash-boards** | **Berichte** |
| **Pentaho DI** | Ja | Ja | Ja | ✘ | ✘ | ✘ | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png |
| **Talend OS** | Ja | Ja | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png |
| **Jasper ETL** | Ja | Ja | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png |
| **Pentaho Mondrian** | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja |
| **Jedox** | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja | Ja |
| **BIRT** | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja | Ja |
| **SQL Power Wabit** | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja | Ja |
| **KNIME** | Ja | Ja | Ja | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja | Ja |
| **RapidMiner** | Ja | Ja | Ja | Ja | Ja | Ja | Ja |
| **Weka** | Ja | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png |  | Ja | https://upload.wikimedia.org/wikipedia/commons/thumb/b/ba/Red_x.svg/15px-Red_x.svg.png | Ja |
| **JasperSoft** | Ja | Ja | Ja | Ja |  | Ja | Ja |
| **Pentaho** | Ja | Ja | Ja | Ja | Ja | Ja | Ja |
| **SpagoBI** | Ja | Ja | Ja | Ja | Ja | Ja | Ja |

Den Anwendungsgebieten entsprechend lassen sich die aufgeführten Open-Source-Programme den Bereichen ETL, OLAP, Data Mining und Berichtwesen zuweisen. Zudem finden sich integrierte BI-Lösungen, die alle aufgeführten Anwendungsgebiete abdecken.

#### ETL-Software

Für die Datenerfassung und Integration im Rahmen eines ETL-Prozesses bieten sich die Open-Source-Programme **Pentaho DI**, **Talend OS** und **Jasper ETL** an.

* **Pentaho DI** [Mehr zu Pentaho DI auf der Anbieterseite] (http://community.pentaho.com/projects/data-integration/)**:** Das auch unter dem Namen Kettle bekannte ETL-Werkzeug Pentaho Data Integration (DI) ist Teil der Pentaho-BI-Suite, kann aber auch unabhängig von den anderen Pentaho-Komponenten als Standalone-Applikation in Data-Warehouse-Architekturen zum Einsatz kommen. Das Datenerfassungs- und Integrations-Tool verfügt über eine grafische Benutzeroberfläche, die auch Anwendern ohne Programmierkenntnisse ermöglicht, ETL-Prozesse zu verwalten. Pentaho DI bietet dazu eine umfangreiche Auswahlpalette diverser Prozess-Bausteine, mit denen sich die einzelnen Arbeitsschritte des ETL-Prozesses definieren lassen. Das Datenintegrationstool unterstützt alle gängigen Datenbanksysteme. Zudem lassen sich Flat-Files wie CSV-, Excel- oder Textdateien als Datenquellen nutzen. Darüber hinaus stellt das Tool Schnittstellen zu proprietären BI-Suiten von SAS oder SAP sowie zu Analyse-Software wie Google Analytics zur Verfügung.
* **Talend OS** [Mehr zu Talend OS auf der Anbieterseite](https://de.talend.com/products/talend-open-studio)**:** Vergleichbar mit Pentaho DI ist das Open-Source-ETL-Werkzeug des Software-Anbieters Talend.Auch Talend Open Studio (OS) ermöglicht Anwendern, Datenerfassungs- und Integrationsprozesse mithilfe parametrisierter Bausteine (sogenannter Jobs) zu definieren. Das Programm bietet Schnittstellen zu allen gängigen Datenquellen und diverse Datentransformationsfunktionen. Ein Map-Editor erlaubt es Anwendern, heterogene Rohdaten in eine vordefinierte Zielstruktur zu übertragen. Wie Pentaho DI kommt auch Talend OS Anwendern ohne Programmierkenntnisse mit einer grafischen Benutzeroberfläche entgegen.
* **Jasper ETL** [Mehr zu Jasper ETL auf der Anbieterseite] (http://community.jaspersoft.com/project/jaspersoft-etl)**:** Jasper ETL ist das Ergebnis einer Kooperation der Softwarehersteller Jaspersoft und Talend. Das ETL-Werkzeug basiert im Wesentlichen auf Talend OS, dem marktführenden Datenintegrations-Tool im Open-Source-Bereich.Ein Einsatz bietet sich vor allem dann an, wenn im Rahmen der DWH-Architektur noch andere BI-Produkte des Anbieters Jaspersoft zum Einsatz kommen.

#### OLAP-Anwendungen

Etablierte OLAP-Werkzeuge unter Open-Source-Lizenz sind **Pentaho Mondrian** und **Jedox**.

* **Pentaho Mondrian** [Mehr zu Mondrian auf pentaho.com] (http://community.pentaho.com/projects/mondrian/)**:** Bei Mondrian handelt es sich um einen java-basierten OLAP-Server. Ursprünglich als eigenständiges Open-Source-Projekt entwickelt, ist Mondrian seit 2006 Teil der Pentaho-BI-Suite. Anwendern steht die Software auch weiterhin als Standalone-Applikation zur Verfügung. Zudem kommt Mondrian in den BI-Lösungen anderer Open-Source-Anbieter wie Jaspersoft zum Einsatz. Anwender profitieren so von einer Bündelung der Open-Source-Ressourcen, die gemeinsame Projekte wie die Mondrian Schema Workbench oder die OLAP4J-Schnittstelle ermöglicht. Das Mondrian-Projekt verfolgt einen relationalen Ansatz (ROLAP). Datenbasis bildet eine relationale Datenbank, deren Tabellen in Stern- oder Schneeflocken-Schemata organisiert sind. Der Zugriff erfolgt in Form mehrdimensionaler Abfragen (MDX), via XML for Analysis (XMLA) oder über die Java-Schnittstelle OLAP4J. Mit der Mondrian Schema Workbench steht Anwendern eine grafische Benutzeroberfläche zur Verfügung. Mondrian Schemata lassen sich bequem am Desktop entwickeln und testen.
* **Jedox** [Mehr zur Jedox auf der Anbieterseite] (http://www.jedox.com/de/produkt/bi-software-testen/jedox-base-download/)**:** Der Softwarehersteller Jedox bietet mit der gleichnamigen BI-Suite eine Komplettlösung für Business-Intelligence- und Performance-Management-Anwendungen an. Zentraler Bestandteilder Software ist ein leistungsstarker In-Memory-OLAP-Server, der über Schnittstellen für Java, PHP, C/C++ oder .NET auch in andere Softwareumgebungen integriert werden kann. Für Anwender im Bereich KMU bietet sich Jedox vor allem aufgrund des Excel-Add-ins an, über das sich der OLAP-Server auch mittels der bekannten Tabellenkalkulationssoftware von Microsoft bedienen lässt. Office-Anwendungen sind in kleinen und mittleren Betrieben weit verbreitet und stellen dort oft die Basis der Datenhaltung dar. Die Excel-Integration reduziert somit den Aufwand für Einarbeitung- und Mitarbeiterschulung.

#### Data Mining

Auch im Bereich Data-Mining stehen Anwendern quelloffene Produkte unter Open-Source-Lizenz zur Verfügung. Das BMWI empfiehlt **KNIME**, **RapidMiner** und **Weka.**

* **KNIME** [Mehr über KNIME auf der Anbieterseite](https://www.knime.org/)**:** KNIME steht für **„**Konstanz Information Miner“, ein Data-Mining-Werkzeug, das an der Universität Konstanz als freie Software entwickelt wird und Anwendern mit einem modularen Pipeline-Konzept neben eigenen Analyseverfahren weitreichende Integrationsmöglichkeiten für diverse Datamining- und Machine-Learning-Algorithmen bietet. Einzelne Schritte der Datenvorverarbeitung (ETL), Modellierung, Analyse und Visualisierung lassen sich über eine grafische Benutzeroberfläche definieren, indem die jeweiligen Bausteine per Drag & Drop in den Arbeitsbereich gezogen und aneinandergereiht werden. Ein kostenloser Download der Software wird von der KNIME.com AG mit Sitz in Zürich angeboten. Bei Bedarf erhalten Anwender dort auch professionelle technische Unterstützung und Beratungsdienstleistungen. Das in Java geschriebene Programm wird als Plug-in für das Programmierwerkzeug Eclipse (IDE) angeboten.
* **RapidMiner** [Mehr zu Rapid Miner auf der Anbieterseite] (https://rapidminer.com/)**:**

Die Analyse-Plattform RapidMiner des gleichnamigen Softwareunternehmens stellt Anwendern eine integrierte Umgebung für maschinelles Lernen, Data-, Text- und Web-Mining, Stimmungs- und Zeitanalysen sowie Vorhersagemodelle in einem Open-Core-Model zur Verfügung. Der Support umfasst alle Stufen des Data-Mining-Prozesses inklusive Datenaufbereitung, Visualisierung, Validierung und Optimierung. Anwendern, denen die kostenlose Community-Version mit nur einem logischen Prozessor und einem Analyseumfang von maximal 10.000 Datensätzen nicht ausreicht, haben die Möglichkeit, ein Upgrade auf eine kostenpflichtige Enterprise-Lizenz zu erwerben. Das Programm ist in Java geschrieben und bietet eine grafische Benutzeroberfläche, mit der sich der Analyse-Workflow bequem per Mausklick definieren und ausführen lässt.

* **Weka** [Mehr zu Weka auf der Website der University of Waikato] (http://www.cs.waikato.ac.nz/ml/weka/)**:** Bei Weka (Waikato Environment for Knowledge Analysis) handelt es sich um ein Open-Source-Projekt der Universität von Waikato, Neuseeland. Das Analyse-Werkzeug bietet Anwendern diverse Algorithmen im Rahmen des maschinellen Lernens. Neben klassischen Data-Mining-Verfahren wie Klassifikation, Assoziation sowie Regressions- oder Cluster-Analysen enthält Weka diverse Bausteine zur Datenvorverarbeitung und Visualisierung. Das in Java geschriebenen Programm bietet eine grafische Benutzeroberfläche. Alle Softwarefeatures lassen sich jedoch auch über die Kommandozeile ausführen. Über eine Java-Schnittstelle lässt sich Weka bei Bedarf in eigene Softwarelösungen integrieren.

#### Berichtwesen

Empfehlenswerte Open-Source-Werkzeuge im Rahmen des Berichtwesens sind **BIRT** und **SQL Power Wabit**. Diese bieten neben klassischen Monats-, Quartals- und Jahresberichten auch Ad-hoc-Funktionen, mit denen sich relevante Informationen in Echtzeit bereitstellen lassen.

* **BIRT** [Mehr zu BIRT auf der Anbieterseite] (http://www.eclipse.org/birt/)**:** BIRT (Business Intelligence and Reporting Tools) ist ein Open-Source-Projekt der gemeinnützigen Eclipse Foundation, das BI-Reporting-Funktionalitäten für Rich Clients und Web-Applikationen bereitstellt. Die Software eignet sich für java-basierte Anwendungen und deckt weite Bereiche der Datenvisualisierung und des Berichtwesens ab. Designs für BIRT-Berichte werden in einer grafischen Benutzeroberfläche erstellt, die auf dem quelloffenen Programmierwerkzeug Eclipse basiert, und als XML-Dateien gespeichert.
* **SQL Power Wabit** [Mehr zu SQL Power Wabit auf der Anbieterseite] (http://www.sqlpower.ca/page/wabit)**:** Mit dem Reporting-Werkzeug SQL Power Wabit erstellen Anwender Berichte auf Basis klassischer Datenbankabfragen. OLAP-Würfel werden nur dann unterstützt, wenn eine Beschreibung der Datenstruktur vorliegt. Das Tool unterstützt Standardberichte, Ad-hoc-Abfragen, benutzerdefinierte Übersichtsseiten und Drill-Down-Operationen im Rahmen des Online Analytical Processings. Mit Funktionalitäten wie einer Drag-&-Drop-Steuerung, der Aktualisierung der Ergebnisberichte in Echtzeit, einer globalen Suchfunktion und einem WYSIWYG-Editor für den Entwurf von Berichten eignet sich SQL Power Wabit auch an Anwender ohne SQL-Kenntnisse. Mit diesem lassen sich umfangreiche Berichte bequem per Mausklick erstellen und bei Bedarf hinsichtlich Schriftart, Farben und Layout individuell anpassen**.**

#### Integrierte BI-Lösungen

Neben den kostenpflichtigen BI-Suiten etablierter Anbieter wie SAP, Oracle, IBM, SAS, HP oder Microsoft gibt es auf dem Open-Source-Markt auch Software-Projekte, die Anwendern Data-Warehousing-Lösungen als integrierte Programmsammlungen zur Verfügung stellen. Zu empfehlen sind **Pentaho CE**, **Jaspersoft** und **SpagoBI**.

* **Pentaho Community Edition (CE)** [Mehr zur Pentaho Community Edition auf community.pentaho.com] (http://community.pentaho.com/)**:** Die Pentaho BI Suite umfasst neben Eigenentwicklungen auch eine Reihe bestehender Open-Source-Projekte, die nach und nach aufgekauft und ins Produktportfolio integriert wurden. Schwerpunkte des Projekts liegen auf der Datenintegration und der Automatisierung von Berichten. Die Programmsammlung umfasst:

* + **Pentaho Business Analytics Platform:** Die **BA Platform** ist eine Web-Applikation, die es Anwendern ermöglicht, alle Informationen in einer zentralen Plattform zusammenzuführen.
  + **Pentaho Data Integration.** Bei Pentaho DI handelt es sich um das oben beschriebene ETL-Tool.
  + **Pentaho Report Designer (PRD):** PRD ist eine Weiterentwicklung des Projekts JFreeReport.Die quelloffene Reporting-Lösung unterstützt diverse Ausgabeformate wie PDF, Excel, HTML, Text, Rich-Text-File, XML und CSV.
  + **Pentaho Marketplace:** Der Marketplaceermöglicht es Anwendern, die Pentaho-Plattform per Mausklick um Plug-ins zu erweitern.
  + **Pentaho Aggregation Designer (PAD):** Mit PADerstellen und optimieren Anwender Datenbankinhalte. Herzstück des Tools bildet der OLAP-Server Mondrian.
  + **Pentaho Schema Workbench (PSW):** Bei PSW handelt es sich um ein grafisches Design-Interface, das es Anwendern ermöglicht, Schemata für Mondrian OLAP-Würfel zu erstellen und zu testen.
  + **Pentaho Metadata Editor (PME):** PME dient der detaillierten Beschreibung der zugrundeliegenden Datenstrukturen mithilfe einer XML-Datei.

Mit Pentaho Enterprise Edition (EE) steht eine kostpflichtige Version der BI-Suite mit erweitertem Funktionsspektrum und professionellem Support zur Verfügung.

* **Jaspersoft** [Mehr zur Community-Version von Jaspersoft] (http://community.jaspersoft.com/)**:** Auch Jaspersoft bietet verschiedene DWH-Anwendungen in einer integrierten BI-Lösung an. Die Programmsammlung umfasst:
  + **JasperReports Server:** Der JasperReports Server ist ein Berichtsserver, der OLAP-Funktionalitäten über einen angepassten Mondrian-Server zur Verfügung stellt.
  + **JasperReports Library:** Zur Erstellung von Berichten stellt Jaspersoft eine Java-Bibliothek zur Verfügung.
  + **Jaspersoft Studio:** Mit Jaspersoft Studio bietet die BI-Suite einen Editor für die Berichterstellung.
  + **Jaspersoft ETL:** Das auf Talend OS basierende ETL-Werkzeug wurde weiter oben bereits beschrieben.
  + **Mobile BI:** Bei Mobile BI handelt es sich um eine native App für iPhone and Android Geräte, die einen mobilen Zugriff auf Berichte und Dashboards zur Verfügung stellt.

Auch Jaspersoft steht mit erweitertem Funktionsspektrum in einer kostenpflichtigen, kommerziellen Version zur Verfügung.

* **SpagoBI** [Mehr zu SpagoBI auf spagobi.org] (https://www.spagobi.org/)**:** Anders als Pentaho und Jaspersoft, die ihre Produkte mit einem dualen Lizenzsystem vermarkten, bietet die Initiative SpagoWorld ihre BI-Suite ausschließlich als Open-Source-Lösung an. Enterprise-Anwender haben jedoch die Möglichkeit, eine professionelle Einrichtung und Anpassung der Software als kostenpflichtige Dienstleistung in Anspruch zu nehmen. Die Programmsammlung umfasst folgende Komponenten:
  + **SpagoBI Server**: Kern der quelloffenen BI-Suite ist der **SpagoBI-**Server, der sämtliche Analysetool- und Funktionalitäten bereitstellt.
  + **SpagoBI Studio**: Bei SpagoBI Studio handelt es sich um eine integrierte Entwicklungsumgebung.
  + **SpagoBI Meta**: SpagoBI Meta bietet Anwendern eine Umgebung für das Metadaten-Management.
  + **SpagoBI SDK**: Mit SpagoBI SDK verfügt die Spago-BI-Suite über eine Integrationsschicht, die es ermöglicht, diverse externe Tools einzubinden: beispielsweise Talend OS (ETL), Jedox oder Mondrian (OLAP), Weka oder R (Datamining) sowie BIRT oder JasperReports Library (Berichtwesen).

#### Datenhaltung

Auch im Bereich der Datenhaltung stehen Anwendern diverse Alternativen zu proprietären Datenbankmanagementsystemen wie Microsoft SQL Server, IBM DB2 oder Lösungen von Oracle und Teradata als quelloffene Software-Projekte zur Verfügung. Als zentraler Datenspeicher bieten sich die relationalen Datenbanksysteme **MySQL [**Mehr zu MySQL auf der Projekt-Seite] (https://www.mysql.de/) und **MariaDB** [Mehr zu MariaDB in Data-Warehouse-Szenarien unter mariadb.com](https://mariadb.com/kb/en/mariadb/data-warehousing-techniques/) oder das objektrelationale DBMS **PostgreSQL** [Mehr zu PostgreSQL unter postgresql.org] (https://www.postgresql.org/) an. Letztgenanntes wird von Pivotal unter dem Namen **Greenplum Database** [Mehr zu Greenplum auf der Projektseite] (http://greenplum.org/) als optimierte Weiterentwicklung speziell für Data-Warehouse-Architekturen unter Open-Source-Lizenz angeboten.

## Fazit: Data-Warehousing im Mittelstand

Data-Warehousing ist im Mittelstand angekommen. Der Markt für BI-Lösungen und Data-Warehouse-Systeme bietet neben kostspieligen Enterprise-Lösungen eine breite Palette brauchbarer Open-Source-Projekte. Für kleine und mittlere Unternehmen sinkt damit vor allem die finanzielle Hürde, die mit einem Einstig Welt der Big-Data-Analysen verbunden ist.

Das BMWI empfehlt Anwendern aus dem Mittelstand, bei der **Einführung von BI-Lösungen** zunächst das Berichtswesen ins Visier zu nehmen. Erste Mehrwerte erzielen Unternehmer durch das Zusammenführen bestehender Daten bereits mit überschaubaren Ausgaben. Treten im Zuge der Auswertung Lücken im Datenbestand zutage, sollte im nächsten Schritt die **Reorganisation der Datenerfassung** mithilfe der hier vorgestellten ETL- oder OLAP-Werkzeuge ins Visier genommen werden. Den Abschluss der Integration einer Data-Warehouse-Architektur in die jeweilige IT-Infrastruktur bilden Data-Mining-Werkzeuge, die durch weitergehende Analysen (z. B. Warenkorbanalysen) **neue Trends und Querverbindungen** aufzeigen können und damit wichtigen Input für strategische Entscheidungen liefern.

Mittelständler, die den Aufbau eines Data-Warehouse ins Auge fassen, sollten von Anfang an auf eine **datenschutzkonforme Umsetzung der jeweiligen BI-Strategie** achten.