# Ein schlankes Data Warehouse dank Big Data

**Dr. Andrea Kennel, InfoPunkt Kennel GmbH**

***Big Data ist die neue Lösung für alle Probleme. Auch im Bereich Data Warehouse ist Big Data eines der großen Themen. Es weckt Träume nach schnellen, günstigen Lösungen, die im Nu neue Kennzahlen ermöglichen und das Business wirklich intelligent machen.***

Die aktuelle Realität sieht oft noch anders aus. Von einem Data Warehouse (DWH) werden viele Kennzahlen mit hoher Qualität verlangt. Die Schwierigkeit liegt darin, die Daten in der vorhandenen Zeit täglich in der gewünschten Qualität zur Verfügung zu stellen. Hier helfen noch mehr Daten mit noch weniger Struktur und schlechter Qualität wenig. Es geht viel eher darum, alle Änderungen und Erweiterungen der Quellen sinnvoll in das bestehende DWH einzubauen.

Der Artikel zeigt zuerst generell die Möglichkeiten von Big Data in einer DWH-Umgebung, einerseits als neue Architektur, aber auch, wie ein bestehendes DWH von Big Data profitieren könnte. Neben den bekannten Ansätzen wird vor allem die Möglichkeit aufgezeigt, wie das DWH bewusst in ein Archiv- und in ein Reporting-DWH unterteilt werden kann. Das Archiv-DWH ist dank Big Data flexibel, das Reporting-DWH durch das Big-Data-Archiv schlank. Dadurch entstehen Projekte, die gleichzeitig dick und schlank sind. So dick, dass alle Daten vorhanden sind, so schlank, dass Reports schnell und hilfreich sind.

## Big Data als Detektiv

In New York gab es in der Kanalisation Probleme mit Fettklumpen, weil Restaurants illegal Frittier-Öl in die Kanalisation entsorgten [1]. Die stichprobenartig vorgenommenen Kontrollen durch Inspektoren fanden zwar einige Übeltäter, aber zu wenige. So ließ man den Computer Auffälligkeiten suchen und kam zu dem Schluss, dass Restaurants, die in der Nähe eines Gullys liegen und auf einen Fett-Abholservice verzichten, mit hoher Wahrscheinlichkeit ihr Öl illegal entsorgen. So konnten die Inspektoren diese gezielt kontrollieren und waren erfolgreich.

Die Möglichkeiten, aus großen Datenmengen unstrukturierter oder wenig strukturierter Daten Informationen zu gewinnen, wird Big Data genannt. Es steht dabei erstmal nur für die große Datenmenge, die oft auf verschiedenen Computern verteilt ist. In klassischen Datenbanken werden Daten für Auswertungen in einer großen Datenbank strukturiert gespeichert und dann ausgewertet. Dabei weiß man, welche Information man aus den Daten gewinnen will. Beispielsweise möchte man die Kunden ermitteln, die mehr als einen gewissen Betrag auf ihrem Bankkonto haben, um diese gezielt für eine Anlageberatung anzuschreiben.

Bei Big Data geht es darum, dass Auffälligkeiten und Regeln nicht bekannt sind. Man lässt über die Daten viele Programme mit mathematischen Modellen laufen, die Auffälligkeiten suchen. So lassen sich Muster in den Daten erkennen, wie der Zusammenhang zwischen der Nähe eines Gullys und dem Verzicht auf einen Abtransport von Öl. Solche Verfahren werden auch Data Mining genannt. Wie in einer Goldmine wird in den Daten gegraben in der Hoffnung, etwas Wertvolles zu finden.

## Big Data erkennt Grippe

Was tut man, wenn die Stimme heiser ist, die Nase läuft und die Stirne heiß wird? Man setzt sich mit einer Tasse warmen Tee an den Computer und googelt. Wenn man dann am nächsten Tag noch dieselben Symptome hat, macht man sich auf den Weg in die Apotheke oder bittet jemanden, Medikamente für einen zu besorgen.

In beiden Fällen hinterlässt man Datenspuren. Einerseits registriert die Apotheke, wenn viele Grippe-Medikamente gekauft werden. Google aber weiß schon früher, dass man nach Heiserkeit, Fieber und anderen Stichworten gesucht hat [2]. So kann mit den Such-Stichwörtern schneller erkannt werden, dass es eine Grippewelle gibt und wo sich diese aktuell ausbreitet. Big Data wird allgemein mit drei „V" beschrieben:

- *Volume*
  Viele Daten
- *Velocity*
  Veränderung zeitkritisch
- *Variety*
  Verschiedenartigkeit

## Viele Daten

Wie viele Daten sind viel und machen ein Volumen aus? Etwas salopp gesagt, kann eine Datenmenge, die sich nicht mehr mit Excel bearbeiten lässt, bereits als groß eingestuft werden. Achtung, die Datenmenge allein genügt aber nicht für Big Data. Viele Datenbank-Lösungen und vor allem große DWH-Lösungen könnten nie mit Excel nachgebaut werden, da sie Terabytes von Daten beinhalten. Trotzdem funktionieren solche Systeme ohne Big Data.

## Veränderung zeitkritisch

Veränderungen in Daten ist auch im DWH-Bereich ein wichtiges Thema. So werden oft Veränderungen in Dimensionen historisiert, also mit dem gesamten Verlauf im DWH gespeichert. So lässt sich jederzeit prüfen, wann sich etwas verändert hat. Je nachdem, was sich ändert, wird dies für das Business relevant, und je schneller man darauf reagieren kann, desto besser.

Wird beispielsweise bei einer Bank ein großer Geldeingang verzeichnet, muss schnell reagiert werden. Es ist zu überprüfen, ob dieser Geldeingang rechtens und unproblematisch ist oder ob die Quelle des

Geldes genauer geprüft werden muss. Hier kann eine Verzögerung von einem Tag, wie in einem DWH üblich, verheerend sein.

Auch die Änderung des Zivilstands kann für eine Bank interessant sein. So ist es für das Marketing wichtig, alle frisch Verheirateten mit geeigneten Finanzprodukten zu bewerben. Hier aber kommt es auf den einen Tag nicht an. Diese Veränderung der Daten ist nicht zeitkritisch.

### Verschiedenartigkeit

Woher stammen die Daten und wie sind diese gespeichert? Eine häufige Anforderung eines DWH ist der sogenannte „Kunden-Flash". Dabei will man alle kundenrelevanten Daten auf einen Blick sehen. Je nach Firma kommen diese Daten aus sehr unterschiedlichen Quellsystemen und sind nicht zwingend aufeinander abgestimmt. Dadurch ist eine Zusammenführung der Daten nicht immer einfach möglich.

Bei dem Beispiel mit den Fettklumpen stammen die Daten nicht nur aus technisch, sondern auch aus organisatorisch verschiedenen Quellen. So kann davon ausgegangen werden, dass diese Systeme nicht aufeinander abgestimmt sind. Im Falle der Grippewelle kommt noch die Schwierigkeit hinzu, dass nicht alle benötigten Daten in Form von Tabellen einer relationalen Datenbank zur Verfügung stehen. Hier ist klar auch Textsuche in wenig strukturierten Daten angesagt.

### Mehrwert von Big Data

Der Mehrwert von Big Data ist im Grunde erst gegeben, wenn alle drei „V" erfüllt sind; wenn also viele Daten aus unterschiedlichen Quellen in unterschiedlicher Form zusammengeführt werden müssen. Konkret, wenn man in einem großen Datenhaufen, der nicht wirklich aufgeräumt ist, Zusammenhänge suchen will.

Da wir nun die drei „V" von Big Data kennen, können wir uns überlegen, ob und wo diese in einem DWH von Nutzen sein können. Schauen wir zunächst an, was das Ziel eines DWH ist. Ein DWH liefert Kennzahlen. Hier die Definition für „Kennzahl" von Wikipedia: „Eine Kennzahl ist eine Maßzahl, die zur Quantifizierung dient und der eine Vorschrift zur quantitativen reproduzierbaren Messung einer Größe oder eines Zustandes oder Vorgangs zugrunde liegt."

Zur Berechnung von Kennzahlen ist das erste „V" oft gegeben. Dass bei Kennzahlen Veränderungen zeitkritisch sind, ist selten der Fall. Das dritte „V" ist für Kennzahlen eigentlich nie gegeben. Per Definition geht es ja darum, dass einer Kennzahl Messpunkte und Größen zugrunde liegen. Diese sind in der Regel in wenigen, klar definierten Quellen mit klaren Strukturen hinterlegt.

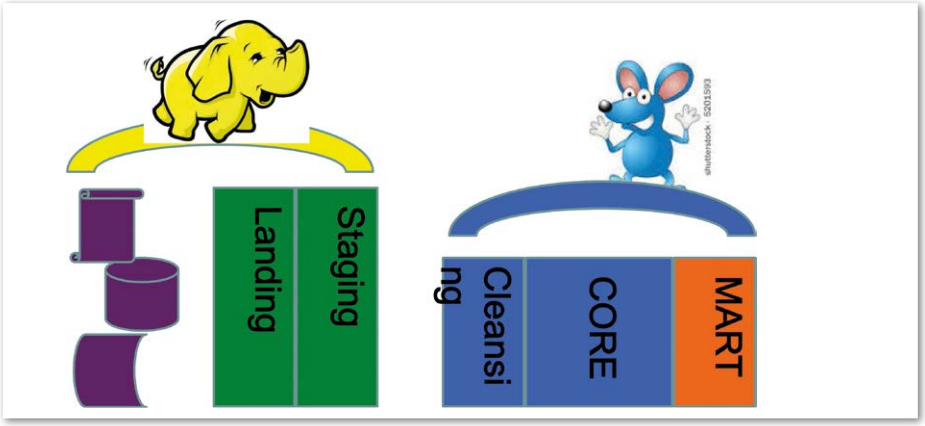


*Abbildung 1: Aufbau eines typischen DWH*

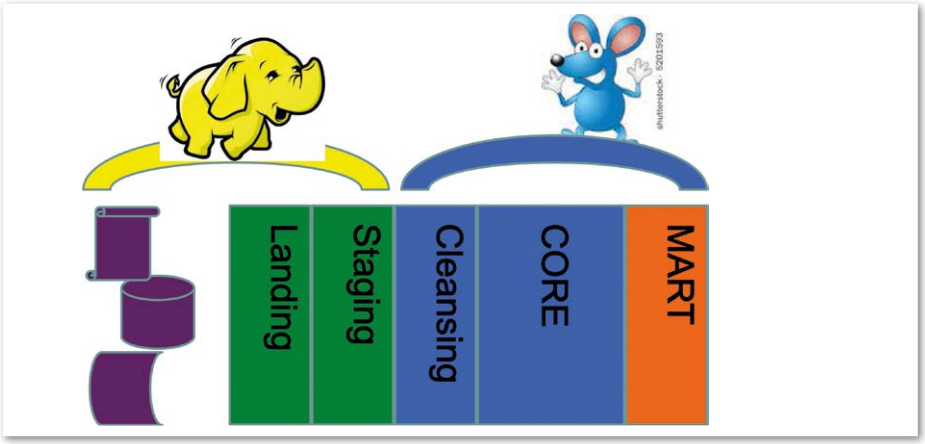


*Abbildung 2: Aufteilung in Archiv und Reporting*

### Vergleich zwischen Dachboden und DWH

Ein typisches DWH hat in der Regel zwei Aufgaben *(siehe Abbildung 1):*

- Archiv
- Reporting

Als Archiv soll ein DWH möglichst alle Daten der Quellen mit allen Veränderungen archivieren. Für das Reporting wird aber normalerweise nur ein Teil der vorhandenen Daten verwendet. Denn nicht alle Daten enthalten interessante Informationen, die man auswerten will. Denn das würde wiederum eine unübersichtliche Informationsflut bedeuten.

Oft stellt sich die Frage, wo die Trennlinie zwischen Archiv und Reporting angesetzt werden kann. Ist das Archiv nur in der Staging oder auch im Core? Werden im Core nur die Daten geladen, die auch zwingend in Mart ausgewertet werden? Sollen alle Daten strukturiert im Core abgelegt werden oder nur diejenigen dort geladen werden, die dann auch für das Reporting benötigt werden?

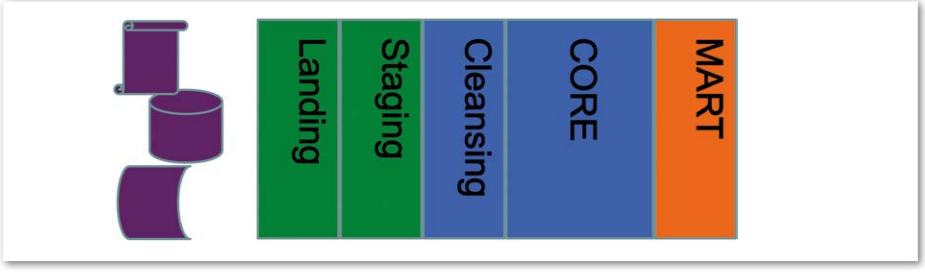


*Abbildung 3: Aufteilung in Archiv und Reporting in zwei DWH*

Die meisten DWH-Projekte ziehen hier keine klare Linie. Normalerweise werden im Core zwar nicht alle Daten geladen, aber doch mehr, als für die Reports nötig sind. Damit wird ein Nachladen vermieden. Bevor die Daten in den Core geladen werden, werden diese in der Cleansing aufgeräumt.

Dieser Vorgang kann mit einem Dachboden verglichen werden. Alles, was wir eventuell einmal wieder brauchen werden, legen wir (in der „Landing") für den Dachboden bereit. Wenn wir das nächste Mal zum Dachboden gehen, nehmen wir all diese Dinge mit und legen sie irgendwo hin. Sobald wir etwas brauchen, beginnt die Suche. Daher ist es sinnvoll, die Dinge, die wir häufig brauchen, auch sauber aufzuräumen und im Dachboden in das vorgesehene Gestell (im „Core") abzulegen.

Beginnen wir ein neues Hobby, so brauchen wir ein neues Gestell, um die entsprechenden Dinge auch sauber aufzubewahren (das ist eine neue Tabelle im „Core"). Wer aber hat Lust und auch den Platz, im Dachboden immer alles sauber aufgeräumt zu haben? Die Dinge, die wir vielleicht nie mehr brauchen, sicherheitshalber aber doch aufbewahren, werden kaum fein säuberlich sortiert, höchstens sehr grob, oder eben da hingestellt, wo diese gerade Platz finden. Denn die Skier passen nicht in ein Büchergestell.

Ein Dachboden ist jedenfalls mit Big Data vergleichbar. Es hat sich über die Jahre so einiges verschieden strukturiert angesammelt. Wenn man aber mal Zeit und Lust zum Suchen hat, so findet man immer interessante Dinge, an die man nicht mehr gedacht hat. Auf dem Dachboden gibt es immer viel zu entdecken.

## Der gelbe Elefant und die blaue Maus

Zurück zum DWH: Für das Reporting müssen die Daten klar und vor allem einheitlich strukturiert sein. Wenn man wissen will, wie viele Paar Skier man hat, will man nicht lange suchen, sondern ganz einfach im Skigestell nachschauen und zählen. Für die Archivierung von Daten lohnt es sich aber nicht immer, diese bereits in einer Form zu strukturieren, die sich für klassische Auswertungen eignet. Somit ist es sinnvoll, nur die „Landing" und „Staging" als Archiv zu betrachten. Diese können und dürfen Daten in einer quellnahen Struktur speichern. Damit können Daten in der „Staging" verschiedenartig und vor allem, je nach Ladevorgang, auch „near-time" vorhanden sein. Daten bis in die „Staging" müssen nicht einheitlich strukturiert sein. Hier ist Big Data sicher hilfreich *(siehe Abbildung 2)*.

Daten im Core sind typischerweise strukturiert, da diese Daten für Reporting genutzt werden. Hier sind klassische, relationale Datenbanken geeignet. Dies ist in der Abbildung mit der blauen Maus dargestellt. Somit können die beiden Aufgaben „Archiv" und „Reporting" in der DWH-Architektur aufgeteilt sein.

Das Archiv enthält, wie erwähnt, verschiedenartige Strukturen. Je nach Quellen kann es interessant sein, die Daten möglichst einfach und schnell in beliebige Strukturen abzulegen. Daher ist dieser Bereich in der Abbildung mit dem gelben Elefanten dargestellt, der auch für Hadoop und damit für Big Data steht.

Konkret können Daten, die als Files angeliefert werden, direkt als Files gespeichert sein. Eine Umwandlung in eine Tabelle entfällt. Wenn dann in der Quelle ein neues Attribut hinzukommt, das für keinen Report notwendig ist, so ändert sich einfach die Struktur der täglich gelieferten Files. Dabei ist im Archiv nichts anzupassen. Es muss nur sichergestellt sein, dass genug Platz zur Verfügung steht.

## Konsequenzen einer Trennung von Elefant und Maus

Die beiden Systeme „gelber Elefant" und „blaue Maus" lassen sich getrennt betrachten. Der gelbe Elefant muss als Archiv alle Daten enthalten und der blauen Maus zur Verfügung stellen. Die blaue Maus braucht jedoch nicht alle Daten aus dem Archiv. In das Reporting-DWH werden nur genau die Daten geladen, die in Reports oder für Ad-hoc-Reporting erforderlich sind.

Diese Möglichkeit besteht nicht erst seit Big Data. Im Jahr 2008 durfte die Autorin bei einem Kunden eine Studie darüber erstellen, wie die Daten archiviert werden können und wie sich das Reporting verbessern lässt. Obwohl es naheliegend war, diese beiden Fragestellungen zu koppeln und ein klassisches DWH bereitzustellen, hat sie sich die Mühe gemacht, die beiden Probleme separat zu betrachten. Das Ergebnis war eine blaue Maus. Die gelbe Elefant war allerdings nicht gelb und auch kein DWH, sondern stand für die Datenquellen selber.

Konkret wurde beschlossen, die Daten weiterhin nicht zusätzlich zu archivieren, sondern alle historischen Daten wie bisher in der Quelle zu halten. Damit war man in der Lage, das DWH in kleinen Schritten aufzubauen. Innerhalb von wenigen Monaten stand die erste schlanke Version des DWH, das die ersten dringenden Reports lieferte.

Bei den Erweiterungen war man in der Lage, das DWH für jedes Release komplett neu aufbauen. So konnte man auf mühsame Daten-Migrationen im DWH verzichten. Auch Korrekturen waren einfach möglich. Das Mapping wurde korrigiert und die Daten mit einem „Full load" neu geladen.

Ein weiterer Vorteil war das Backup-Recovery-Konzept, das man sehr vereinfachen konnte. Ein Backup oder teure Datenspiegelung waren nicht nötig. Die Recovery wurde mit jedem Release getestet, denn das hieß: Alle Strukturen löschen, neu anlegen und die Daten neu laden.

Dass alle Daten in der Quelle auch historisch vorhanden sind, ist sicher die Ausnahme. Daher lohnt sich die Überlegung, wie ein schlankes Reporting-DWH möglich wird, wenn die Quelle nicht als Archiv dienen kann. Ja, genau: Die Antwort lautet „gelber Elefant". Hier ist Big Data nicht nur ein Schlagwort, sondern klar eine neue Möglichkeit, Daten von verschiedenen Quellen systematisch zu archivieren *(siehe Abbildung 3)*.

Ist das Archiv-DWH als Big Data umgesetzt und das Reporting nur mit den wirklich nötigen Daten gefüllt, so ergeben sich folgende Vorteile:

- Kein Cleansing für Archiv erforderlich
- Keine Strukturänderung für Archiv
- Neue Kennzahlen können rückwirkend aus Archiv ermittelt werden
- Reporting lässt sich ab Archiv immer neu laden

## Fazit

Ein DWH, das für Kennzahlen-Reporting benutzt wird, muss klar strukturierte Daten beinhalten. Big Data hat andere Stärken, die auf verteilten, unstrukturierten oder unterschiedlich strukturierten Daten aufbauen. Big Data kann jedoch für die Archivierung von Daten und somit für ein Archiv-DWH interessant sein. Daher ist eine Unterteilung eines DWH in Archiv und Reporting wichtig. Dank eines großen Archivs mit Big Data kann sich das Reporting-DWH auf die Daten konzentrieren, die für Reporting erforderlich sind, und bleibt damit schlank und einfach.

## Weitere Informationen

[1] GEO eBook: Big Data, Die neue Intelligenz des Menschen, S. 4-6
[2] Ginsburg, Jeremy et al., Detecting influenza epidemics using search engine query data, http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html
[3] GEO eBook: Big Data, Die neue Intelligenz des Menschen, S.14

*Dr. Andrea Kennel*
*andrea@infokennel.ch*