Daten und Unternehmen – Data Warehousing

Marlene Johanna Festag

Seminararbeit im Interdisziplinären Lehrangebot
des Instituts für Informatik

Leitung: Prof. Hans-Gert Gräbe, Ken Pierre Kleemann

`http://bis.informatik.uni-leipzig.de/de/Lehre/Graebe/Inter`

Leipzig, 29.09.2017

Abbildungsverzeichnis

Inhaltsverzeichnis

1. Einleitung

Sprechen wir heutzutage vom *digitalen Wandel*, betrachten wir zwar einerseits Lebensbedingungen in Zeiten von ständiger Vernetzung, jedoch sind diese auch untrennbar mit einer sich drastisch verändernden Arbeitswelt verbunden.
Hierbei haben wir bereits verschiedene Entwicklungsphasen mit entsprechenden Anforderungen überwunden, von den Anfängen der Computerisierung über die Etablierung dieser bis hin zu unserer heutigen weltumspannenden Vernetzung.
In dieser Arbeit soll es dabei insbesondere um das Data Warehouse gehen, das in seiner Idee nicht nur technisch interessant ist, sondern auch die Schwelle zu einer neuen Mentalität in Unternehmen markiert, in der Informationen in der Form von Daten als Wettbewerbsvorteil und wertvolle Währung gesehen werden können.
Ein solides Datenmanagement kann sich auf Kostenersparnissen, dem besseren Erfüllen von Kundenerwartungen und der Produktivität der Mitarbeiter niederschlagen.
Somit werden die grundlegenden Fragen in dieser Arbeit thematisiert:
Wozu ein Data Warehouse? Welche Bedeutung hat die Instanz innerhalb der Unternehmensstruktur? Und vor allem: ist dieser Entwicklungsschritt bereits durch Big Data beziehungsweise Data Streams überholt, und es wird auf verschiedene Implementierungen eingegangen.

1.Unternehmen im Wandel

Um den digitalen Wandel und damit auch das Data Warehouse zu verstehen, beschäftigen wir uns zunächst einmal mit der Entwicklung die all dem zugrunde liegt - der sogenannten *Computerisierung*.
Wurde im 2. Weltkrieg der Computer vor allem für militärische sowie vereinzelt für wissenschaftlich und technische Zwecke genutzt, begann in den 50er und 60er Jahren schließlich die kommerzielle Verwendung. Jedoch erst seit den 80er Jahren verbreiteten sich PC und Home-Computer, da besonders die seit 1984 eingeführten grafischen Benutzeroberflächen sowie die Erfindung der Computermaus die Verwendung revolutionierten (Eberle 1991).
In dieser Zeitspanne findet sich auch das erste Mal der Begriff "Data Warehouse" durch Barry Delvin, obwohl ähnliche Konzepte bereits seit den 60er-Jahren in Form von "Business Intelligence" bekannt waren. Anfangs noch das Spielzeug weniger, sorgten sinkende Preise und leistungsstärkere Modelle dafür, dass 1991 fast jeder Betrieb mit über 20 Mitarbeitern computerisiert wurde, und selbst Kleinbetriebe die Anschaffung machten (ebd. 1991).
Die Grundlage für die digitale Revolution war geschaffen: traditionelle Berufe verschwanden zu diesem Zeitpunkt durch computerisierte Prozesse oder verändern sich stark, andererseits botenComputer nun die Möglichkeit, große Mengen an Daten schnell und effektiv zu sichten und auszuwerten, für die vorher noch Karteikarten mit Kundeninformationen beispielsweise hätten durchsucht werden müssen.
Schließlich rückten die Organisation der Arbeitsabläufe immer mehr ins Interesse der Unternehmen und der Bedarf für das Errichten von IT-Strukturen wuchs (vgl. ebd. 1991).
Diese Entwicklungsstufe, welche auch als *Konnektivität* bezeichnet wird, löst den isolierten Gebrauch von technischen Geräten durch den Versuch der Widerspiegelung der Unternehmensstruktur in einem verbundenen, unternehmensinternen Netzwerk ab.
Doch zum modernen Unternehmen führt noch ein weiter Weg - im Laufe der Zeit sorgen Sensoren, Mikrochips und bessere Netzwerktechnik für eine Echtzeiterfassung des Fertigungsprozesses. Doch auch hier stellt sich nun die Frage: wer hat Zugriff auf welche Daten? Kontextuelle Einordnung und Wirkungszusammenhänge müssen untersucht werden

um letztendlich die gewonnen Informationen für aussagefähige Prognosen nutzen zu können. (Schuh 2017)
Haben wir nun die Computerisierung, Echtzeiterfassung, Transparenz und Prognostik-möglichkeiten gegeben, so sind wir beim Status Quo eines modernen Unternehmens angekommen. Nun können wir uns mit der Struktur eines solchen beschäftigen in Bezug auf die Verwendung eines Data Warehouse.

2. Operative und dispositive Systeme

Für ein besseres Verständnis dafür, inwiefern die meisten IT-Strukturen funktionieren, spielt auch eine wichtige Trennung eine Rolle: operative (oder OLAP für "Online Transactional Processing") und dispositive Systeme (OLAP für "Online Analytical Processing") . Entscheidungsunterstützende oder dispositive Systeme dienen analytischen und prozess-optimierenden Zwecken und sind bereits seit Beginn der Computerisierung bekannt, lediglich in unterschiedlichen Formen:
Management Information Systems (MIS), Decision Support Systems (DSS), Executive Information Systems (EIS), später Data Warehouses (DW) und schließlich Business Intelligence (BI)-Lösungen. Diese haben verschiedene Kernideen, bedienen jedoch alle den Bedarf der Unternehmen nach verlässlicher Speicherung wichtiger Informationen . Business Intelligence beschreibt hierbei eine Sammlung an Anwendungen und Technologien zur Verarbeitung entscheidungsunterstützender Daten, wozu auch ein Data Warehouse gehören kann.
Die operativen Systeme wiederum beschäftigen sich mit dem Tagesgeschäft, also dem Anlegen, Lesen, Ändern und Löschen von Daten pro Transaktion (vgl. Humm/Wietek 2005). Oftmals sind die operativen Systeme auf lokale Bereiche im Unternehmen begrenzt und stützen sich auf die Informationen aus den dispositiven Instanzen (vgl. Gabriel 2016).

3. Das Data Warehouse und seine Struktur

Da die operativen Systeme nur für kurzfristige Speicherung ausgelegt sind, um mit der Datenlast das System nicht auszubremsen oder gar überfordern, können alte Daten hier nicht verbleiben. . Dennoch sind historische Daten wichtig um aussagekräftige Informationen zu erhalten und Prognosen zu stellen.

Eine Lösung für dieses Problem stellt hier das Data Warehouse dar.

Dies wird durch folgende Merkmale definiert:

a) Themenorientierung im Bezug auf spezifische Kernbereiche, z.B. Kundendaten
b) Vereinheitlichung der diversen Informationen aus den operativen Systemen. Ziel ist ein konsistenter Bestand der anhand von Indizes abgesucht werden kann. Dabei hilft die Organisation in Spalten bzw. Tabellenform.
c) Zeitorientierung,da in der Analyseschicht vor allem gewisse Zeiträume wie Wochen, Monate oder Jahre betrachtet werden. Somit müssen die Daten durch einen Zeitbezug identifizierbar bleiben, beispielsweise organisiert in Jahr/Monat/Kalenderwoche. Jedoch werden heute oftmals neuere Implementierungen bis hin zu “Realtime Data Warehouse Architekturen” verwendet, da bestimmte Branchen wie die Telekommunikationsbranche sofortige aktuelle Daten statt Turnusauswertungen benötigen (Gluckowski 2012).
d) Beständigkeit. Um über lange Zeiträume hinweg verlässlich Daten zu archivieren, benötigt das Warehouse entsprechende Speichertechniken, die gleichzeitig die für Abfragen benötigte Zeit in Grenzen halten und nur in Ausnahmefällen das Löschen oder Abändern zulassen. Hierzu gehört auch die strenge Trennung des Data-warehouses von den Datenquellen sowie der Analyse - einerseits bleiben die Daten objektiv, andererseits können die Bestände auch getrennt von dem jeweiligen Unternehmen verwendet werden, selbst wenn das Unternehmen selbst nicht mehr existiert oder beispielsweise durch technische Defekte im operativen Sektor lahmgelegt ist. Auch die Unabhängigkeit vom Benutzer trägt zur Beständigkeit bei, da Daten so schwerer zurückgehalten oder manipuliert werden können, sondern

stattdessen ein ganzheitliches Bild der Unternehmenslage unverfälscht gespeichert wird (vgl. ebd. 2012).

Weiterhin hat sich bezüglich des schnellen Zugriffs um die Sinnhaftigkeit eines Data Warehouses gewährleisten zu können ein System der *Data-Marts* etabliert (Humm/Wietek 2005). Zur Gewährleistung der Effektivität verwendet man untergeordnete Marts bzw Datenpools, die auf eine Abteilung und deren spezifische Daten ausgerichtet sind. Hierbei können sie in verschiedenen Formen der Abhängigkeit vom Warehouse sowie in Hybridform existieren.

In folgendem Beispiel werden die Daten aus den operativen Systemen in das Data Warehouse geladen, worin unter anderem in Metadaten, Rohdaten etc unterteilt wird.
Entsprechend werden dann Umsatz und Marketing Daten in zwei untergeordneten Pools einsortiert.

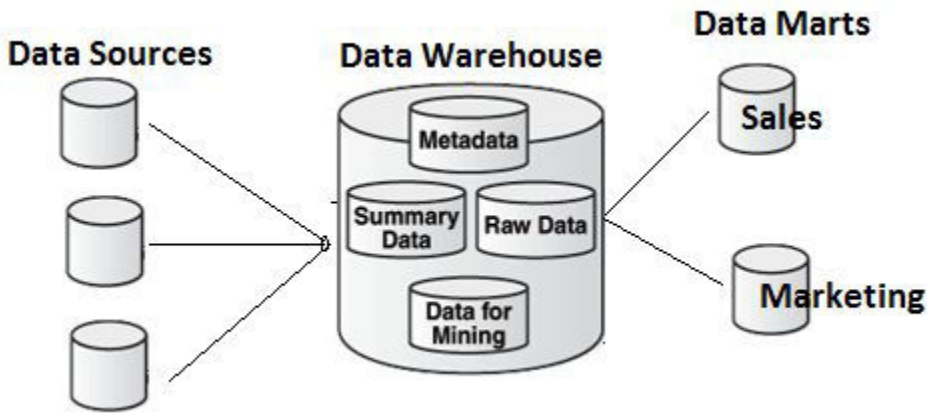

Abb. 2: Beispiel Verhältnis Quelle, Warehouse und Data Marts

4.Bewerten von Daten

Gerade wenn es darum geht, das Data warehouse effektiv zu gestalten spielt es auch eine wichtige Rolle, Informationen eine entsprechende Wichtigkeit zuzuweisen um die unterschiedliche Relevanz zu markieren. Hierbei unterscheidet man

Kritische Daten, welche beispielsweise für die wichtigen Geschäftsprozesse genutzt werden, sowie Bankverbindungen, Verträge, Bestelleingänge und Kundendaten. Diese bilden die primären Daten des Warehouses, da besonders diese auf Langzeitspeicherung angelegt sein sollten.
Eine weitere Kategorie sind *performance Daten* bezüglich der operativen Systeme und des Tagesgeschäfts, die relevant für die Steuerung und Planung des Unternehmens sind, hier ist vor allem eine mittelfristige Speicherung im Data Warehouse vorgesehen.
Nicht-kritische und *sensible Daten* wiederum sind in der Regel entweder leicht wiederherzustellen oder können durch vergleichbare Informationen ersetzt werden.

Neben dieser Klassifizierung hat sich weiterhin in der Praxis ein Tag bezüglich des Lebenszyklus bewährt:
“verwendet, "analysiert", "archiviert" und "gelöscht". Als "verwendet" gelten in operativen Systemen gehaltene Informationen, werden sie nicht mehr geändert, jedoch noch für die Auswertung in dispositiven Systemen verwendet, befinden sie sich im Lebenszyklus "analysiert". Haben sie den Status der Speicherung ins Warehouse komplett vollzogen gelten sie wiederum als archiviert.
Nicht-kritische und Performance Daten werden zur Optimierung des Speicherplatzes, nachdem die entsprechenden Analysezwecke erfüllt wurden, gelöscht (vgl. Liebhart 2014).

5. Das Data Warehouse in der Unternehmensarchitektur

Ein Beispiel, wie eine Unternehmensstruktur mit Data Warehouse aussehen kann, sieht folgendermaßen aus:

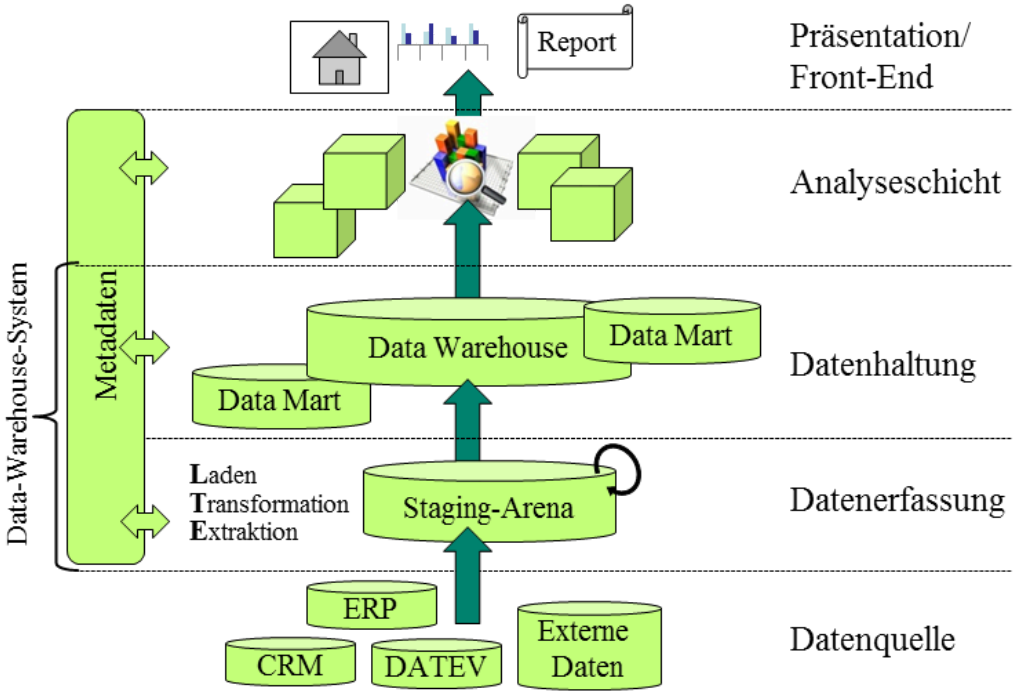

Abb.1: IT-Struktur eines Unternehmens mit Data-Warehouse

Von unten beginnend haben verschiedene Abteilungen entsprechende operative Systeme. So befasst sich hier im Beispiel die ERP mit der Ressourcenplanung, CRM mit den Kundenbeziehungen und DATEV mit Steuer und Management. Zusammen mit externen Daten werden diese nun in der Staging Area vereinheitlicht. Dieser Schritt dient der Informationsintegration und die Daten werden zunächst aus den operativen Systemen extrahiert, in ein dem Data-warehouse dienliches format transformiert und schließlich in selbiges geladen.

Dort werden diese nun sinnvoll archiviert, sodass die Analyseschicht mithilfe verschiedener Methoden wertvolle Informationen aus der Lagerung ziehen kann. Diese Methoden vereint man oftmals unter dem Begriff “Data-Mining”, jedoch ist zu beachten, dass dies kein homogener Begriff, sondern mehr als Schirmterm zu sehen ist (Liebhart 2014).

Letztendlich können die gezogenen Erkenntnisse im front end präsentiert werden und dem

Management fällt die Aufgabe zu, die entsprechenden Konsequenzen zu ziehen.
Die hier gezeigten Metadaten stellen dabei Verweise auf Daten dar.

6. Variationen der Business Intelligence

Da jedes Unternehmen individuelle Strukturen und Ansprüche hat,
ist es kaum verwunderlich, dass auch die IT-Strukturen diese Vielfalt wiederspiegeln. Und so gibt es auch bezüglich der Business Intelligence diverse Variationen mit verschiedenen Vor- und Nachteilen.
Einige davon werden hierbei näher beleuchtet.

6.1 Das konventionelle, strukturierte Data-Warehouse

Ein konventionelles Data-Warehouse besteht aus organisierten Aufzeichnungen in Form von Säulen oder Tabellen welche mit einem Index versehen sind (vgl. Fulton 2013).
Die Datenverarbeitung bzw. das Lesen geschieht mittels SQL, einer auf relationaler Algebra basierter Datenbanksprache. Für die Aufbereitung in Richtung Anwender gibt es wiederum die bereits vorgestellten Data Marts. Für übliche Zwecke, wie sie in einem produzierendem Unternehmen mit typischen Bereichen wie Controlling und Berichtswesen anfallen, sind beide Herangehensweisen geeignet.
Relationale Datenbanken bieten durch sogenanntes Joining die Möglichkeit, jederzeit Daten miteinander zu verbinden und können so auch sofort auf neue Anforderungen eingehen. Sie ermöglichen dadurch Real-Time-Datawarehouses (vgl. Welker 2015).

6.2 Der unstrukturierte Data-Store

Was nun mit unstrukturierten Informationen? Semistrukturierte Daten wie E-Mails besitzen zwar einen gewissen Aufbau, allerdings keine Metadaten, Textdaten in natürlicher Sprache oder Tonaufzeichnungen bezeichnet man wiederum als völlig unstrukturierte Daten. Diese werden in einem Datenpool mit umfassendem Speichervolumen zwischengelagert und heute meist durch eine sogenannte “big data engine” verwaltet (Fulton 2013). Beliebt ist hierbei das open source Framework Hadoop, welches auf dem MapReduce Algorithmus von Google basiert und so die Datenlast des vom Anwender ungesichteten Materials auf mehreren

Rechnern parallel verarbeitet und reduziert wird (ebd. 2013).

6.3. Cloud-basierte Speicherung

Man spricht hierbei auch von Everything as a Service (EaaS), da IT-Strukturen unterschiedlicher Art von externen Anbietern genutzt werden.
So kann auch Speicherplatz von Anbietern wie Amazon oder Rackspace nach Bedarf geleased werden. Der Kunde kann Ressourcen eigenständig reservieren oder freigeben, berechnet werden letztendlich nur die genutzten Reserven, nicht benötigte können dann von einem anderen Kunden verwendet werden . Statt eigenes Personal für die Verwaltung eines Data-Warehouses im IT-Bereich zu beschäftigen, ist in den Kosten für die Cloud die Wartung bereits enthalten, so bietet sich dieses Modell vor allem für kleine, flexibel arbeitende Firmen an. Zeiten hohes Speicherbedarfs wie Monats oder Jahresabschlüsse sind einfach durch das Anfordern weiterer Reserven zu lösen.
Verschiedene Cloud Typen, von privaten Clouds, welche nur eine einzige Firma nutzt, über die Community Cloud die einen eingeschränkten Firmenkreis bedient, bishin zu public Clouds die auch von Privatpersonen verwendet werden können, ist der Flexibilitätsaspekt das Hauptargument dieser Art des BI (vgl. Fehling).

6.4. Data-Streams/DSMS

Unternehmen, die in überwiegendem Maße mit unstrukturierten Daten arbeiten, benötigen eine interaktive, flexible und dynamische Verarbeitung.
Um dem Nutzer Informationen zu stellen, ohne konkret danach zu suchen in einer sowohl reaktiven als auch proaktiven Art, bilden Data Streams eine Alternative zu traditionellen dispositiven BI Systemen und verlaufen eng verbunden mit dem Big Data Trend.
Ein wichtiger Faktor in dieser Entwicklung ist auch die Überholung des traditionellen Unternehmens durch “enterprise 2.0.” Unternehmen, die weniger in festen lokalen Abteilungen als in kollaborativen Kreisen zusammenarbeiten (Aufaure 2016).
Technisch funktioniert hier das Data Stream System folgendermaßen:
Data Stream Management Systems (DSMS) vollziehen kontinuierlich Anfragen. Die Datenelemente werden eingespeist, verbleiben jedoch nur eine begrenzte Zeit im Speicher.

Diese Architektur soll verhindern, das System zu überlasten, welches sonst mit der Archivierung und Analyse weder zeitlich noch in Bezug auf tatsächlichen Speicherplatz hinreichend schnell und effektiv wäre.
Innerhalb dieser begrenzten Speicherzeit extrahieren diverse Tools die Informationen, die für den Nutzer tatsächliche Qualität haben mithilfe eines komplexen Abfrage(Query) Tools.
Der Direktzugriff ist im Gegensatz zum Data-Warehouse nur sequentiell möglich, während die Anfrage kontinuierlich den Stream in Echtzeit durchläuft. Im Warehouse dagegen wird nur eine einmalige Abfrage gemacht, die den momentanen Wissensstand der Datenbank durchforstet (vgl. Aufaure 2016).

7. Fazit

Business Intelligence befindet sich momentan in einem Wandel, insbesondere durch die immer weiter steigende Relevanz und potentieller Anteil unstrukturierter Daten.
Von einer völligen Ablösung des Data Warehouse Modells durch Big Data und reine Echtzeitanalyse in Streams ist aber zur Zeit noch nicht zu sprechen.
Insbesondere zeigt die kurz vorgestellte DSMS Architektur, dass sich unser Anspruch an Daten in einem Umbruch befindet. Die seit den 80er Jahren entwickelte Systematik von Design und Ausführung der Datenbanken ist bei der global exponentiell steigenden Last an Informationen an einen Punkt gelangt, der neuartige Speicher- und Verarbeitungskonzepte erfordert. Gerade im produzierenden Gewerbe mit lokal begrenzten Abteilungen ist das Data Warehouse jedoch weiterhin aktuell und sinnvoll, insbesondere durch die neueren Implementierungen, wie Real-Time und relationierte Architekturen.
Sollten jedoch mehr und mehr Unternehmen zukünftig nicht mehr in der Form lokaler Niederlassungen, sondern ineinem globalen Netzwerk von aufgabenbedingten Zusammenschlüssen, gerade im Personalsektor, existieren, werden besonders Cloud-Speicherung, Hybrid Data-Warehouses und Streams in einer entsprechend dynamischen Struktur an immenser Bedeutung gewinnen.

Literaturverzeichnis

Eberle, Thomas (1991): *Computerisierung*, Luzern: Verl. Schweizer Lexikon.

Mucksch, Harry et al. (2000): *Das Data Warehouse-Konzept als Basis einer unternehmensweiten Informationslogistik,* 4. Auflage, Wiesbaden: Gabler , S. 3 – 80.

Schuh, Günther et al. (2017): *Industrie 4.0 Maturity Index. Die digitale Transformation von Unternehmen gestalten*, München: Herbert Utz Verlag.

Internetverzeichnis

Aufaure, Marie et al. (2016): *From Business Intelligence to semantic data stream management,* in URL: http://www.sciencedirect.com/science/article/pii/S0167739X15003635 (Zugriff: 11.09.2017).

Fehling, Christoph/Leymann, Frank: *Cloud Computing*, in: Springer Gabler Verlag (Hrsg.): Gabler Wirtschaftslexikon, in URL: http://wirtschaftslexikon.gabler.de/Archiv/1020864/cloud-computing-v9.html (Zugriff: 07.09.2017).

Fulton, Scott M. (2013): *Data Warehousing Options - Understanding Data Warehousing."* , in URL: http://www.tomsitpro.com/articles/data_governance-big_data-business_analytics-shadow_it-hadoop,2-549-4.html (Zugriff: 11.09.2017).

Gabriel, Roland (2016): *Operatives Informationssystem*, in URL: http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/uebergreifendes/Kontext-und-Grundlagen/Informationssystem/operatives-informationssystem (Zugriff 12.09.17).

Gluchowski, Peter (2012): *Data Warehouse*, in URL: http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/daten-wissen/Business-Intelligence/Data-Warehouse (Zugriff: 04.09.2017).

Humm, Bernhard,; Wietek, Frank (2005): *Architektur von Data Warehouses und Business*

Intelligence Systemen, in URL: https://www.fbi.h-da.de/fileadmin/personal/b.humm/Publikationen/Humm__Wietek_-_Architektur_DW__Informatik-Spektrum_2005-01_.pdf (Zugriff: 11.09.2017).

Liebhart, Daniel (2014): *die vier Säulen guten Datenmanagements*, in URL: https://www.computerwoche.de/a/die-vier-saeulen-guten-datenmanagements,2555587 (Zugriff 23.9.2017).