

Informationsgenerierung und -Speicherung

André Csillaghy & Nicky Hochmuth, FS 2011

soleil.i4ds.ch/~csillag/teaching/igs

Inhalt

2.1	EINLEITUNG	5
2.1.1	<i>Wieso ein DWH?</i>	5
2.1.2	<i>OLAP und OLTP</i>	9
2.1.3	<i>Drei komplementäre Themenbereiche</i>	10
2.1.4	<i>Herausforderungen beim Data Warehousing</i>	11
2.1.5	<i>Extrahieren, transformieren, laden (ETL)</i>	11
2.2	DAS MULTIDIMENSIONALE DATENMODELL	13
2.3	MOLAP vs. ROLAP.....	13
2.4	OLAP DESIGN: STAR SCHEMA	14
2.5	HIERARCHIEN VON DIMENSIONEN.....	14
2.6	OLAP ABFRAGEN UND TYPISCHE OPERATIONEN AUF DWHS.....	15
2.7	OLAP MIT SQL ABFRAGEN.....	16
2.8	DER CUBE OPERATOR.....	16
2.8.1	<i>CUBE Beispiel mit Oracle SQL</i>	16
2.9	WINDOW QUERIES	18
2.10	NEUE (SQL:2003) AGGREGATE FUNKTIONEN.....	19
2.11	INDIZES FÜR OLAP	20
2.11.1	<i>Bitmap Index</i>	20
2.11.2	<i>Join Index</i>	20
2.11.3	<i>Bitmap Join Index</i>	21
2.12	VIEWS IN DWHS	21
2.12.1	<i>View Materialisation (Vorbereitung)</i>	22
2.12.2	<i>View Materialisation und Indizes</i>	23
2.12.3	<i>Probleme mit View Materialisation</i>	24
2.13	NACH MATERIALISATION: TOP N ABFRAGEN, INTERAKTIVE ABFRAGEN	24
2.13.1	<i>Top N Abfragen</i>	24
2.13.2	<i>Interaktive Abfragen</i>	24
2.14	ZUSAMMENFASSUNG.....	24
3.1	EINLEITUNG	26
3.1.1	<i>Was bringt Data Mining?</i>	26
3.1.2	<i>Die Kette des Data Minings</i>	27
3.1.3	<i>Wieso benutzen Leute Data Mining?</i>	27
3.2	SCHRITTE DES DATA MININGS	28
3.2.1	<i>Knowledge Discovery = Kenntnisbeschaffung</i>	28
3.2.2	<i>Schritte des Data Minings detailliert</i>	28
3.2.3	<i>Beispiel einer Anwendung: Digital Sky Survey</i>	28
3.3	DATA MINING AN OBJEKTMEGEN – MARKET BASKET ANALYSIS	29
3.3.1	<i>Häufige Objektmengen zählen</i>	29
3.3.2	<i>Ein Algorithmus, um häufige Objektmengen zu identifizieren</i>	29
3.3.3	<i>Iceberg Abfragen</i>	30
3.3.4	<i>Regeln suchen</i>	31
3.3.5	<i>Ein Algorithmus, um Regeln zu suchen</i>	31
3.3.6	<i>Assoziationsregeln und ISA Hierarchien</i>	32
3.3.7	<i>Verallgemeinerung der Association Rules</i>	32
3.3.8	<i>Anwendung von Assoziationsregeln</i>	33

3.3.9	<i>Sequentielle Patterns</i>	34
3.4	DATA MINING MODELLE	34
3.4.1	<i>Typen von Daten</i>	35
3.4.2	<i>Typen von Variablen</i>	35
3.5	ÜBERWACHTES LERNEN – SUPERVISED LEARNING	35
3.5.1	<i>Klassifikationsregeln und Regressionsregeln</i>	36
3.6	UNÜBERWACHTES LERNEN: CLUSTERING	37
3.6.1	<i>Clustering: Unüberwachtes Lernen</i>	38
3.6.2	<i>Messung der Ähnlichkeit</i>	40
3.6.3	<i>K-Means</i>	41
3.7	ÜBERWACHTES LERNEN: ENTSCHEIDUNGSBÄUME [DECISION TREES].....	43
3.7.1	<i>Generierung des Baumes</i>	44
4.1	EINLEITUNG: INDIZIERUNG FÜR TEXTSUCHE.....	46
4.2	INVERTIERTE DATEIEN	47
4.3	SIGNATUR DATEIEN [SIGNATURE FILES]	47
4.3.1	<i>Signatur-Dateien: Abfrageevaluation</i>	48
4.4	VECTOR SPACE MODEL	48
4.4.1	<i>Term Frequency – Inverse Document Frequency</i>	49
4.4.2	<i>Länge normalisieren</i>	50
4.4.3	<i>Ähnlichkeit messen</i>	50
4.4.4	<i>Precision und recall</i>	51
5.1	HITS.....	52
5.1.1	<i>HITS Algorithmus</i>	53
5.2	GOOGLE UND DER PAGERANK ALGORITHMUS	53
5.2.1	<i>PageRank Algorithmus</i>	54
5.2.2	<i>Probleme</i>	55
5.2.3	<i>Die Elemente eines Web Search Suche</i>	56
5.3	GOOGLE IN ZAHLEN	57
5.4	GOOGLE ARCHITEKTUR: PRINZIPIEN	58
5.4.1	<i>Wie wird eine Abfrage auf Google ausgewertet?</i>	58

1 Vorbemerkung – Literatur – Web Links

Die Notizen, die sich in diesem Dokument befinden, sind aus meinem Kurs abgeleitet. Sie haben keinen Anspruch, vollständig zu sein. Im Gegenteil. Sie sollen nur als Begleitung der Dokumente gelten, auf welcher sie basieren, eine Art Anhaltspunkte, aus welchen man in weitere Dokumente vollständige Information finden kann. Insbesondere sind die Folgenden Dokumente Pflichtlektüren:

Data Warehouses: Ramakrishnan and Gehrke, Chapter 22

Data Mining: Ramakrishnan & Gehrke, Chapter 23

Modern Information Retrieval:

Sehr viel Informaion ist auch auf dem Netz zu finden. Hier einige Links, die ich nützlich finde und gebraucht habe, um den Kurs vorzubereiten:

2 Data Warehousing

2.1 Einleitung

Unternehmen wollen die Möglichkeiten ihrer IT-Infrastruktur ausnutzen, um aktuelle und historische Daten gründlicher analysieren zu können, als dies von Hand möglich wäre. Unternehmen wollen *Mustern in den Daten* identifizieren, die damit sie helfen, strategisch sinnvolle Entscheidungen treffen zu können (sofern das überhaupt möglich ist). Somit werden in einem Data Warehouse Daten aus verschiedene Bereiche des Unternehmens gesammelt.

Mit Data Warehouses (DWH) will man also sehr grosse Datenmengen analysieren. Die Analyse geschieht meistens

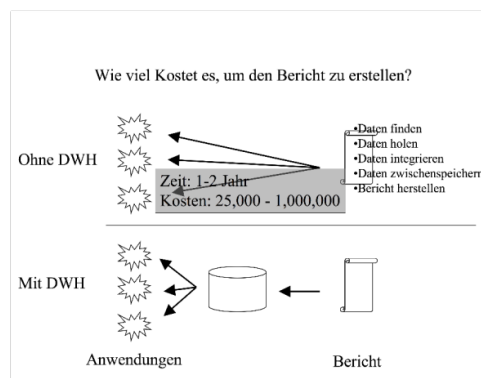
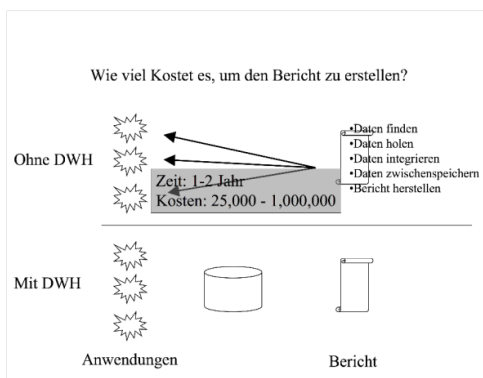
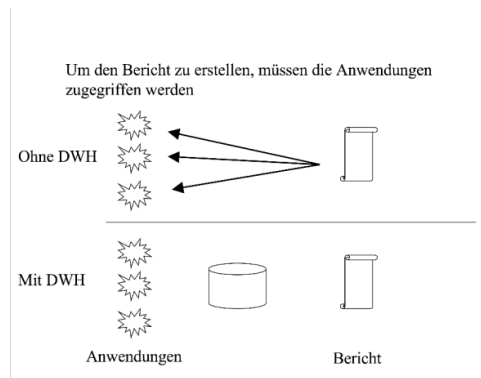
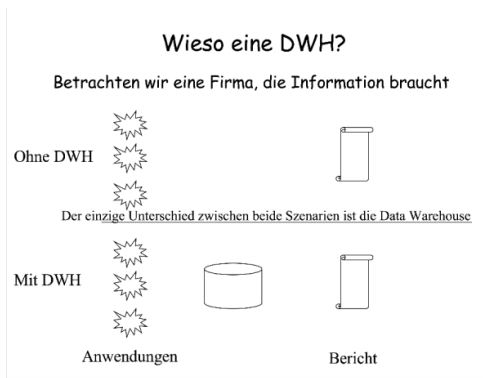
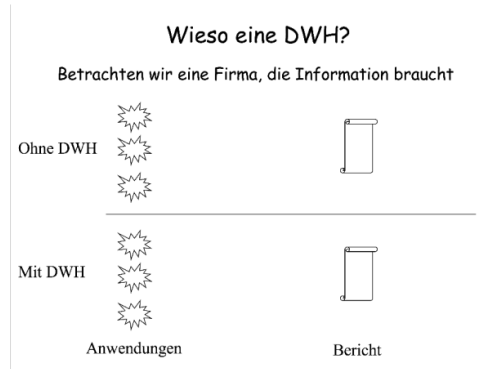
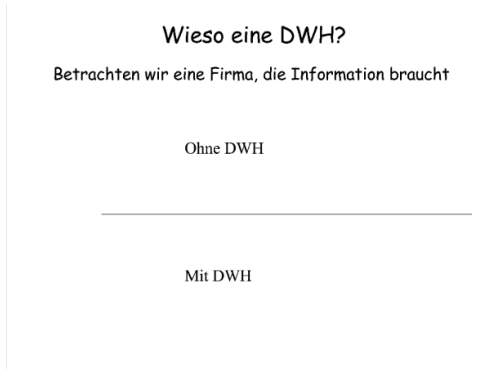
- Interaktiv;
- Entdeckungsartig [explorative];
- Komplex.

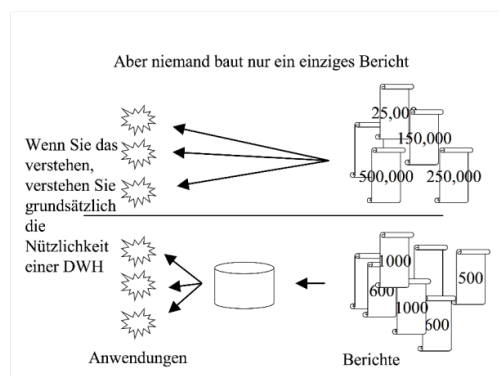
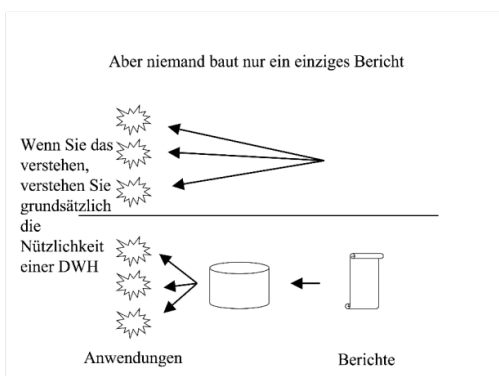
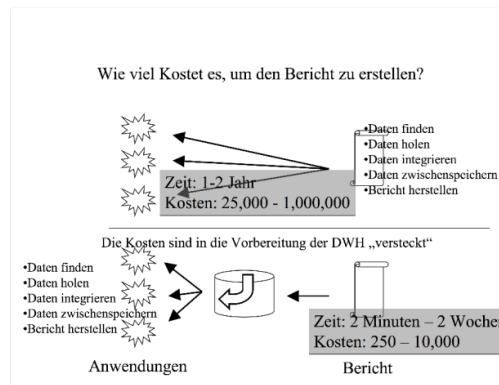
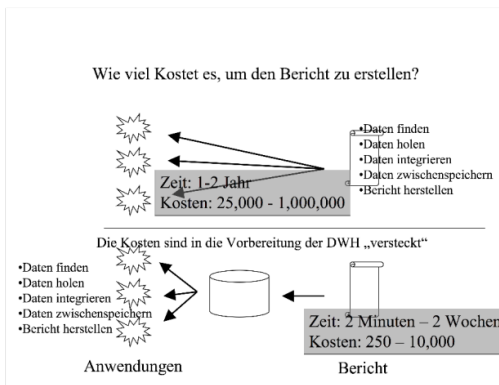
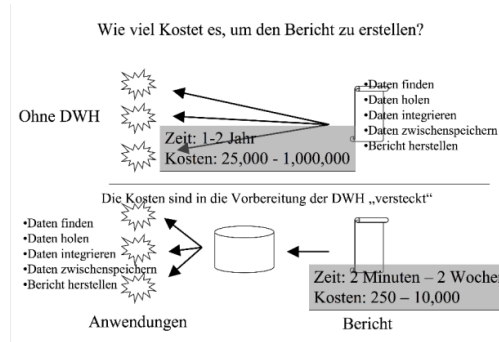
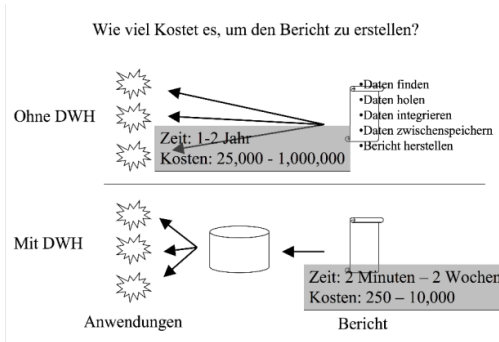
Daten in einem DWH sind mehr oder weniger statisch. Einmal gesammelte Daten werden üblicherweise nie mehr geändert. Man betrachtet sie als ein *snapshot* des Unternehmens.

Aktuelle Unternehmen im Bereich DWH/OLAP sind: Greenplum, Aster, Cloudera, ParAccel, Vertica. Alle fokussieren ihr Business auf dem Management von sehr grossen Datenmengen.

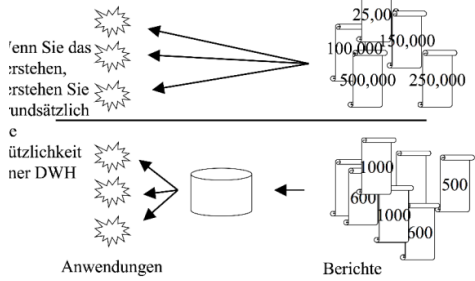
2.1.1 Wieso ein DWH?

Die Motivation, ein DWH einzusetzen, kann mit der bildlichen Beschreibung, die in Abbildung 1 angegeben ist, erklärt werden.



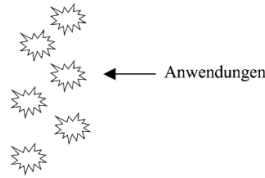


Der Gewinn wird durch die Wiederbenutzung der Berichte realisiert



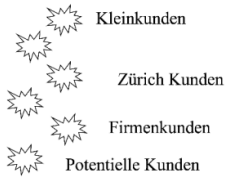
Aber es gibt noch weitere Gründe für eine DWH

- Betrachten wir die Anwendungen, die eine Firma über Zeit kreiert oder gekauft hat:



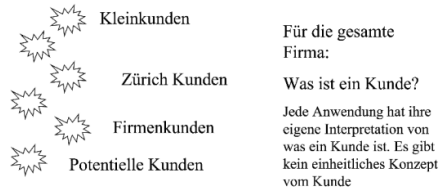
Aber es gibt noch weitere Gründe für eine DWH

- Betrachten wir die Anwendungen, die eine Firma über Zeit kreiert oder gekauft hat:

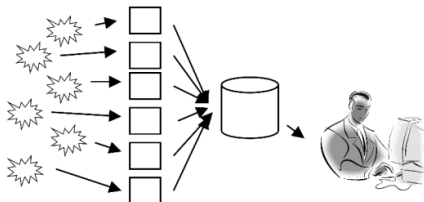


Aber es gibt noch weitere Gründe für eine DWH

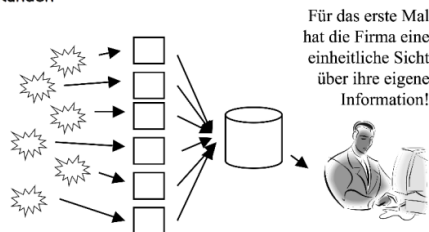
- Betrachten wir die Anwendungen, die eine Firma über Zeit kreiert oder gekauft hat:



- Daten werden aus der einzelnen Anwendungen geholt und in der Data Warehouse integriert
- Mit der Integration werden Daten vereinheitlicht und stellen dann eine einheitliche Sicht über dem Begriff „Kunden“



- Daten werden aus der einzelnen Anwendungen geholt und in der Data Warehouse integriert
- Mit der Integration werden Daten vereinheitlicht und stellen dann eine einheitliche Sicht über dem Begriff „Kunden“



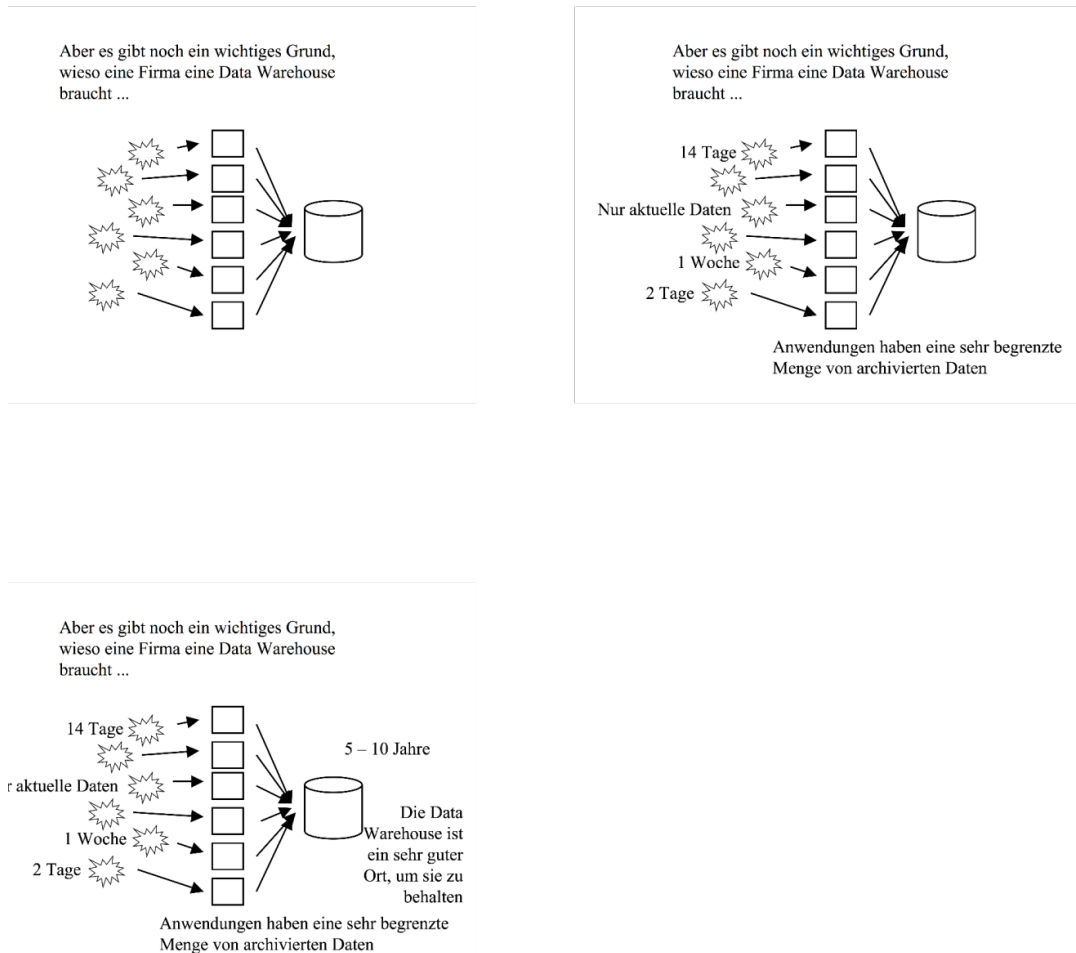


Abbildung 1: Motivation für ein Data Warehouse

2.1.2 OLAP und OLTP

Daten eines DWH stehen in Kontrast zu Daten in einer operativen Datenbank. In einem DWH geht es ausschliesslich um die Analyse der Daten. Operative Datenbanken hingegen behandeln Daten Transaktionsorientiert. Man spricht deshalb von

- **Online Analytical Processing (OLAP)** in DWHs, gegenüber
- **Online Transaction Processing (OLTP)** in traditionellen operativen Datenbanken.

Eine typische OLAP Abfrage ist: „gib mir den Verkauf pro Monat und pro Filiale an“ während eine OLTP Abfrage wäre „gib mir den aktuellen Stand des Kontos.“

Bei DWHs geht es um **wenige** Abfragen, die sehr komplex sind und recht **viel Zeit** in Anspruch nehmen. Bei OLTP geht es um viele, relativ kurze Update-Transaktionen, die wir aus der relationalen DBMS-Welt kennen. Tabelle 1 fasst die Unterschiede zusammen.

Tabelle 1: Unterschiede zwischen OLTP und OLAP

OLTP	OLAP
Transaktionsorientiert	Analysierorientiert
Anwendungsorientiert	Themenorientiert
Abfrageresultate detailliert	Abfrageresultate zusammengefasst
Zeitlich genau	Zeitlich ungenau (snapshots)
Für die Administration	Für Managers
Wird ständig modifiziert (z.B. Jede Millisekunde)	Wird nicht oder selten (im Vergleich mit operativen Datenbanken) modifiziert (z.B. jeden Tag)
Läuft repetitiv	Läuft heuristisch
Untersuchte Prozesse sind a priori verstanden	Untersuchte Prozesse sind a priori nicht verstanden
Performanz-sensitiv	Nicht Performanz-sensitiv
Zugriff Datensatz-orientiert	Zugriff Mengenorientiert

2.1.3 Drei komplementäre Themenbereiche

Bei der Anwendung einer DWH werden drei verbundene Themenbereiche betrachtet:

- **Data Warehousing** bezeichnet die Integration [Engl: consolidation] von Daten aus verschiedenen Quellen in einem einzigen, sehr grossen Archiv:

- Laden von Daten, periodisches Aktualisieren/Synchronisieren aus anderen Systemen;
- Semantische Integration.

- **OLAP:**

- Komplexe SQL-Abfragen und Views;
- Abfragen basieren auf Spreadsheet-Stil. Operationen werden auf „multidimensionalen“ Sichten der Daten durchgeführt;
- Interaktive und online Abfragen, durchgeführt von Endbenutzern.

- **Data Mining:** Entdeckungsartige Suche nach interessanten Trends und Anomalien

Zuerst konzentrieren wir uns auf Data Warehousing und OLAP und werden später auf Data Mining eingehen, da DWH und OLAP meistens die Grundsteine für Data Mining Operationen darstellen. OLAP kann prinzipiell als ein primitives Data Mining betrachtet werden. Data Warehousing kann mit dem Design des Datenmodells assoziiert werden, welches dann für die OLAP Abfragen angewendet wird.

2.1.4 Herausforderungen beim Data Warehousing

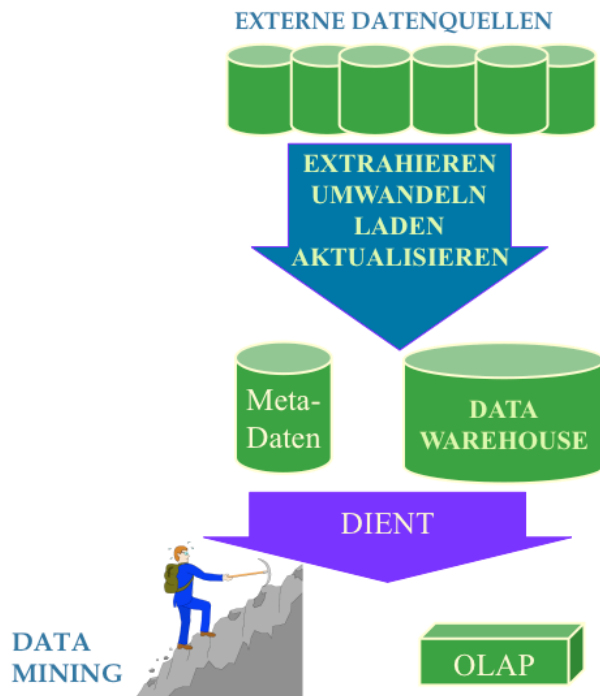


Abbildung 2: Workflow bei Data Mining

Daten werden aus einer operativen Datenbank (oder allgemeiner: aus externen Datenquellen) integriert. Mit Integration bezeichnen wir den Prozess, um die Daten in einem einheitlichen Format bzw. Schema zu bringen. Oft werden sie noch mit zusammengefasster Information ergänzt (Metadaten).

Die heutige Größe von Data Warehouses liegen heute (2011) oft in Größen von Petabytes (z.B. ebay >10TB Teradata Database). Bei solchen Datenmengen stossen „traditionelle“ Datenbanksysteme an ihre Grenzen.

Interaktive Reaktionszeiten sind für Abfragen nötig, da die Daten „on the fly“ von den Benutzern (speziell Manager, die aus Prinzip keine Zeit haben) analysiert werden.

2.1.5 Extrahieren, transformieren, laden (ETL)

Ein DWH erfordert die integrierte Sammlung von unterschiedlich strukturierten Daten aus verschiedenen Quellen. Dies ist bei weitem keine triviale Aufgabe: Die Integration der Daten erfordert Entscheide, die sich im gesamten Design widerspiegeln. Das Prozess der Integration von heterogenen Daten aus (operativen) Umgebungen wird mit ETL bezeichnet:

- **EXTRACT:** Daten werden aus externen, operativen Datenbanken ausgelesen und vorbereitet [cleaned], um einerseits Fehler zu minimieren und andererseits um fehlende Informationen zu ergänzen.
- **TRANSFORM:** Es geht da vor allem um semantischen Integration: Wenn man Daten aus verschiedenen Quellen verarbeitet, muss man unangepasste Informationen eliminieren (z.B.

unterschiedliche Währungen). Die Transformation kann über die Definition einer Sicht [CREATE VIEW] erfolgen.

- **LOAD**: Laden, Aktualisieren, Entfernen [purge]:

- Daten müssen regelmässig geladen, periodisch aktualisiert, und veraltete Daten entfernt werden. Das Laden kann z.B. mit Materialisation der in der „Transform“-Stufe erstellten View realisiert werden. Die Materialisation wird aber in einer anderen Datenbank gespeichert als diejenige, die ihre Definition enthält. Zusätzliche Vorbereitungen sind nötig: Sortieren, Generierung von zusammenfassenden Information (Monats-Summen, usw.), Partitionierung, Generierung von Indizes. Bei Datenmengen von mehreren Terabytes kann das Einfügen von Daten in das DWH sehr lange dauern (siehe Übung 1).
- Aktualisieren: Ähnlich wie Asynchronous Replication in operativen DBMS.

Für alle Informationen im DWH müssen die Datenquelle, Ladezeit und andere Information registriert werden, so dass später rekonstruiert werden kann, welche Daten im DWH überhaupt abgelegt sind! (Stichwort: Metadata Repository)