

Data-Warehouse-Technologien

Prof. Dr.-Ing. Kai-Uwe Sattler¹ Prof. Dr. Gunter Saake²
Dr. Veit Köppen²

¹TU Ilmenau
FG Datenbanken & Informationssysteme

²Universität Magdeburg
Institut für Technische und Betriebliche Informationssysteme

Letzte Änderung: 11.10.2011

Teil II

Data-Warehouse-Architektur

Data-Warehouse-Architektur

- 1 Anforderungen
- 2 Referenzarchitektur
- 3 Phasen des Data Warehousing
- 4 Komponenten

Anforderungen des Data Warehousing

- Unabhängigkeit zwischen Datenquellen und Analysesystemen (bzgl. Verfügbarkeit, Belastung, laufender Änderungen)
- Dauerhafte Bereitstellung integrierter und abgeleiteter Daten (Persistenz)
- Mehrfachverwendbarkeit der bereitgestellten Daten
- Möglichkeit der Durchführung prinzipiell beliebiger Auswertungen

Anforderungen des Data Warehousing

- Unterstützung individueller Sichten (z.B. bzgl. Zeithorizont, Domäne und Struktur)
- Erweiterbarkeit (z.B. Integration neuer Quellen)
- Automatisierung der Abläufe
- Eindeutigkeit über Datenstrukturen, Zugriffsberechtigungen und Prozesse
- Ausrichtung am Zweck: Analyse der Daten

12 OLAP-Regeln nach Codd

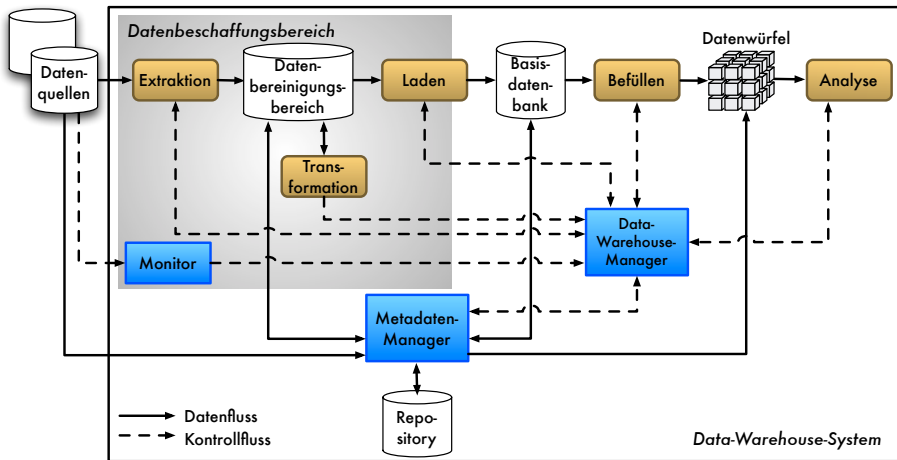
- Multidimensionale, konzeptionelle Sicht
- Transparenz
- Zugriffsmöglichkeit
- Performanz
- Skalierbarkeit
- Generische Dimensionalität
- Dynamische Handhabung dünnbesetzter multidimensionaler Strukturen
- Mehrbenutzerbetrieb
- Uneingeschränkte Operationen
- Intuitive Benutzeroberfläche
- Flexibles Reporting
- Beliebig viele Dimensionen und Aggregationsebenen

FASMI

Fast Analysis on Shared Multidimensional Information

- Kurze Antwortzeiten (im Mittel unter fünf Sekunden)
- Einfache und flexible Möglichkeiten von Auswertungen
- Heterogene Benutzer mit unterschiedlichen Rechten
- Multidimensionalität ist wichtiges Kriterium
- Fragen nach Anzahl benötigter Dimensionen und Wertebereiche zugehöriger Attribute

Referenzarchitektur



Phasen des Data Warehousing

- 1 Überwachung der Quellen auf Änderungen durch Monitore
- 2 Kopieren der relevanten Daten mittels Extraktion in temporären Datenbereinigungsbereich
- 3 Transformation der Daten im Datenbereinigungsbereich (Bereinigung, Integration)
- 4 Kopieren der Daten in integrierte Basisdatenbank als Grundlage für verschiedene Analysen
- 5 Befüllen der Datenwürfel (Datenbanken für Analysezwecke)
- 6 Analyse: Operationen auf Daten des DW

Basisdatenbank und Datenwürfel stellen das Data Warehouse dar.

Data-Warehouse-Manager

- Zentrale Komponente eines DW-Systems
- Initiierung, Steuerung und Überwachung der einzelnen Prozesse (Ablaufsteuerung)
- Initiierung des Datenbeschaffungsprozesses
 - ▶ In regelmäßigen Zeitabständen (jede Nacht, am Wochenende etc.): Starten der Extraktion von Daten aus Quellen und Übertragung in Datenbereinigungsbereich
 - ▶ Bei Änderung einer Quelle: Start der entsprechenden Extraktionskomponente
 - ▶ Auf explizites Verlangen des Administrators
 - ▶ Push vs. Pull Strategie,
 - ▶ Aktualität ist Anforderung für Analyseaufgaben

Data-Warehouse-Manager

- Nach Auslösen des Ladeprozesses:
 - ▶ Überwachung der weiteren Schritte (Bereinigung, Integration etc.)
 - ▶ Koordination der Reihenfolge der Verarbeitung
- Fehlerfall
 - ▶ Dokumentation von Fehlern
 - ▶ Wiederanlaufmechanismen
- Zugriff auf Metadaten aus dem Repository
 - ▶ Steuerung des Ablaufs
 - ▶ Parameter der Komponenten

Datenquellen

- Lieferanten der Daten für das Data Warehouse
 - ▶ Gehören nicht direkt zum DW
 - ▶ Können intern (Unternehmen) oder extern (z.B. staatl. Einrichtung) sein
 - ▶ Heterogen bzgl. Struktur, Inhalt und Schnittstellen (Datenbanken, Dateien)
 - ▶ Auswahl der Quellen und Qualität der Daten von besonderer Bedeutung
- Faktoren für Auswahl
 - ▶ Zweck des DW
 - ▶ Qualität der Quelldaten
 - ▶ Verfügbarkeit (rechtlich, sozial, technisch)
 - ▶ Preis für Erwerb der Daten (speziell bei externen Quellen)

Datenquellen: Klassifikation

- Herkunft: intern, extern
- Zeit: aktuell, historisch
- Nutzungsebene: Primärdaten, Metadaten
- Inhalt: Zahl, Zeichenkette, Grafik, Referenz, Dokument
- Darstellung: numerisch, alphanumerisch, BLOB
- Sprache und Zeichensatz
- Vertraulichkeitsgrad

Datenquellen: Qualitätsforderungen

- Konsistenz (Widerspruchsfreiheit),
- Korrektheit (Übereinstimmung mit Realität),
- Vollständigkeit (z.B. Abwesenheit von fehlenden Werten oder Attributen),
- Zuverlässigkeit (z.B. Vertrauen in die Datenquellen),
- Genauigkeit (z.B. Anzahl der Nachkommastellen),
- Granularität (z.B. tagesgenaue Daten),
- Zeitnähe (Wann wurde die letzte Änderung durchgeführt vs. Auftreten der Datenänderung),
- Relevanz (Wie wichtig sind die Daten?),
- ...

Datenquellen: Qualitätsforderungen (2)

- Zuverlässigkeit (Nachvollziehbarkeit der Entstehung, Vertrauenswürdigkeit des Lieferanten),
- Verständlichkeit (inhaltlich und technisch / strukturell für jeweilige Zielgruppe),
- Verwendbarkeit (geeignetes Format, Zweckdienlichkeit),
- Einheitlichkeit (Datenformat),
- Eindeutigkeit (Interpretierbarkeit) und
- Schlüsselintegrität (Schlüssel und Referenzen)

Monitore

- Aufgabe:
 - ▶ Entdeckung von Datenmanipulationen in einer Datenquelle
- Strategien:
 - ▶ Trigger-basiert
 - ★ Aktive Datenbankmechanismen
 - Auslösen von Triggern bei Datenänderungen
 - Kopieren der geänderten Tupel in anderen Bereich
 - ▶ Replikationsbasiert
 - ★ Nutzung von Replikationsmechanismen zur Übertragung geänderter Daten

Monitore (2)

- Strategien (fortg.):
 - ▶ Log-basiert
 - ★ Analyse von Transaktions-Log-Dateien der DBMS zur Erkennung von Änderungen
 - ▶ Zeitstempelbasiert
 - ★ Zuordnung eines Zeitstempel zu Tupeln
 - ★ Aktualisierung bei Änderungen
 - ★ Identifizierung von Änderungen seit der letzten Extraktion durch Zeitvergleich
 - ▶ Snapshot-basiert
 - ★ Periodisches Kopieren des Datenbestandes in Datei (Snapshot)
 - ★ Vergleich von Snapshots zur Identifizierung von Änderungen

Datenbereinigungsbereich

- Aufgabe:
 - ▶ Zentrale Datenhaltungskomponente des Datenbeschaffungsbereichs (engl. staging area)
 - ▶ Temporärer Zwischenspeicher zur Integration
- Nutzung:
 - ▶ Ausführung der Transformationen (Bereinigung, Integration, etc.) direkt auf Zwischenspeicher
 - ▶ Laden der transformierten Daten in DW bzw. Basisdatenbank erst nach erfolgreichem Abschluss der Transformation
- Vorteile:
 - ▶ Keine Beeinflussung der Quellen oder des DW
 - ▶ Keine Übernahme fehlerbehafteter Daten

Extraktionskomponente

- Aufgabe: Übertragung von Daten aus Quellen in Arbeitsbereich
- Funktion: abhängig von Monitoring-Strategie
 - ▶ Periodisch
 - ▶ Auf Anfrage
 - ▶ Ereignisgesteuert (z.B. bei Erreichen einer definierten Anzahl von Änderungen)
 - ▶ Sofortige Extraktion
- Realisierung:
 - ▶ Nutzung von Standardschnittstellen (z.B. ODBC, JDBC)
 - ▶ Ausnahmebehandlung zur Fortsetzung im Fehlerfall

Transformationskomponente

- Vorbereitung und Anpassung der Daten für das Laden
 - ▶ Inhaltlich: Daten-/Instanzintegration und Bereinigung
 - ▶ Strukturell: Schemaintegration
- Überführung aller Daten in ein einheitliches Format
 - ▶ Datentypen,
 - ▶ Datumsangaben,
 - ▶ Maßeinheiten,
 - ▶ Kodierungen, etc.
- Beseitigung von Verunreinigungen (engl. Data Cleaning bzw. Data Cleansing)
 - ▶ Fehlerhafte oder fehlende Werte,
 - ▶ Redundanzen,
 - ▶ Veraltete Werte.

Transformationskomponente (2)

- Data Scrubbing:
 - ▶ Ausnutzung von domänenspezifischen Wissen (z.B. Geschäftsregeln) zum Erkennen von Verunreinigungen
 - ▶ Beispiel: Erkennen von Redundanzen
- Data Auditing:
 - ▶ Anwendung von Data-Mining-Verfahren zum Aufdecken von Regeln
 - ▶ Aufspüren von Abweichungen

Ladekomponente

- Aufgabe:
 - ▶ Übertragung der bereinigten und aufbereiteten (z.B. aggregierten) Daten in die Basisdatenbank bzw. das DW
- Besonderheiten:
 - ▶ Nutzung spezieller Ladewerkzeuge (z.B. SQL*Loader von Oracle)
→ Bulk-Laden
 - ▶ Historisierung: Änderung in Quellen dürfen DW-Daten nicht überschreiben, stattdessen zusätzliches Abspeichern
- Ladevorgang:
 - ▶ Online: Basisdatenbank bzw. DW steht weiterhin zur Verfügung
 - ▶ Offline: stehen nicht zur Verfügung (Zeitfenster: nachts, Wochenende)

Basisdatenbank

- Aufgabe:

- ▶ Integrierte Datenbasis für verschiedene Analysen
→ unabhängig von konkreten Analysen, d.h. noch keine Aggregationen
- ▶ Versorgung des DW mit bereinigten Daten (u.U. durch Verdichtung)

- Anmerkungen:

- ▶ Wird in der Praxis oft weggelassen
- ▶ Entspricht Operational Data Store (ODS) nach Inmon

Datenwürfel

- Aufgabe: Datenbanken für Analysezwecke (relational oder multidimensional)
- Orientieren sich in Struktur an Analysebedürfnissen
- Basis: DBMS
- Besonderheiten:
 - ▶ Unterstützung des Ladeprozesses
 - ★ Schnelles Laden großer Datenmengen
→ Massenslader (engl. bulk loader) unter Umgehung von Mehrbenutzerkoordination und Konsistenzprüfung
 - ▶ Unterstützung des Analyseprozesses
 - ★ Effiziente Anfrageverarbeitung (Indexstrukturen, Caching)
 - ★ Multidimensionales Datenmodell (z.B. über OLE DB for OLAP)

Data Warehouse

Im engeren Sinn:
Basisdatenbank und Datenwürfel stellen das Data Warehouse dar.

- Im weiteren Sinn stellen die Data Marts ebenfalls Komponenten des Data Warehouses dar.

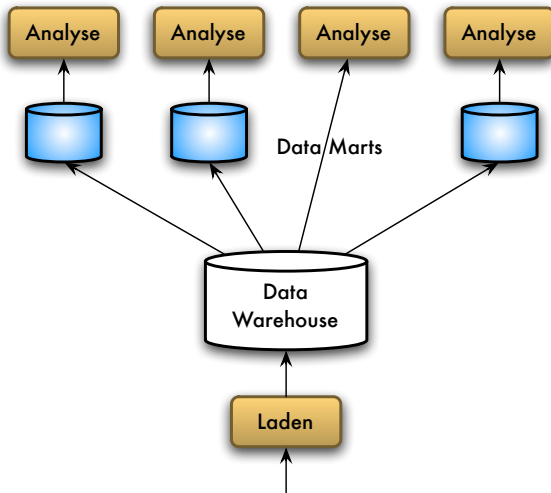
Data Marts

- Aufgabe:
 - ▶ Bereitstellung einer inhaltlich beschränkten Sicht auf das DW (z.B. für Abteilung)
- Gründe:
 - ▶ Eigenständigkeit, Datenschutz, Lastverteilung, Datenvolumen, etc.
- Realisierung:
 - ▶ Verteilung der DW-Daten
- Formen:
 - ▶ Abhängige Data Marts
 - ▶ Unabhängige Data Marts

Abhängige Data Marts

- Verteilung des Datenbestandes nach
 - ▶ Integration und Bereinigung (Basisdatenbank) und
 - ▶ Organisation entsprechend der Analysebedürfnisse (Datenwürfel)
- „Nabe- und Speiche“-Architektur (engl. hub and spoke)
- Data Mart:
 - ▶ Nur Extrakt (inkl. Aggregation) des Data Warehouse
 - ▶ Keine Bereinigung oder Normierung
- Analysen auf Data Mart konsistent zu Analysen auf DW
- Einfache Realisierung:
 - ▶ Replikations- oder Sichtmechanismen von DBMS

„Nabe- und Speiche“-Architektur



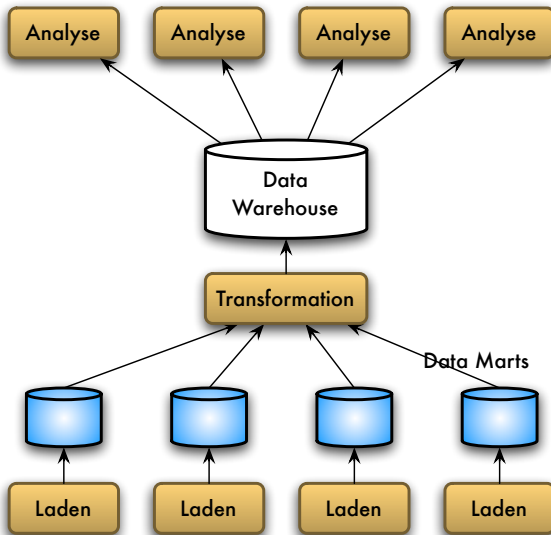
Abhängige Data Marts: Extraktbildung

- Strukturelle Extrakte
 - ▶ Beschränkung auf Teile des Schemas
 - ▶ Bsp.: nur bestimmte Kennzahlen oder Dimensionen
- Inhaltliche Extrakte
 - ▶ Inhaltliche Beschränkung
 - ▶ Bsp.: nur bestimmte Filialen oder das letzte Jahresergebnis
- Aggregierte Extrakte
 - ▶ Verringerung der Granularität
 - ▶ Bsp.: Beschränkung auf Monatsergebnisse

Unabhängige Data Marts

- Unabhängig voneinander entstandene „kleine“ Data Warehouses (z.B. von einzelnen Organisationen)
- Nachträgliche Integration und Transformation
- Probleme:
 - ▶ Unterschiedliche Analysesichten (Data Mart, globales Data Warehouse)
 - ▶ Konsistenz der Analysen aufgrund zusätzlicher Transformation

Unabhängige Data Marts



Analysewerkzeuge

- Engl. Business Intelligence Tools
- Aufgabe:
 - ▶ Präsentation der gesammelten Daten
 - ▶ interaktive Navigation
 - ▶ Analysemöglichkeiten
- Analyse:
 - ▶ Einfache arithm. Operationen (z.B. Aggregation) ... komplexe statistische Untersuchungen (z.B. Data Mining)
 - ▶ Aufbereitung der Ergebnisse für Weiterverarbeitung bzw. Weitergabe

Analysewerkzeuge: Darstellung

● Tabellen

- ▶ Pivot-Tabellen := Kreuztabellen
Merkmalsausprägungen in Zeilen- und Spaltenüberschrift
- ▶ Analyse durch Vertauschen von Zeilen und Spalten
- ▶ Veränderung von Tabellendimensionen
- ▶ Schachtelung von Tabellendimensionen

Umsatz		Bier	Rotwein	Summe
2009	Sachsen-Anhalt	45	32	77
	Thüringen	52	21	73
	Summe	97	53	150
2010	Sachsen-Anhalt	60	37	97
	Thüringen	58	20	78
	Summe	118	57	175

● Graphiken

- ▶ Bildliche Darstellung großer Datenmengen
- ▶ Netz-, Punkt-, Oberflächengraphen

● Text- und Multimedia-Elemente

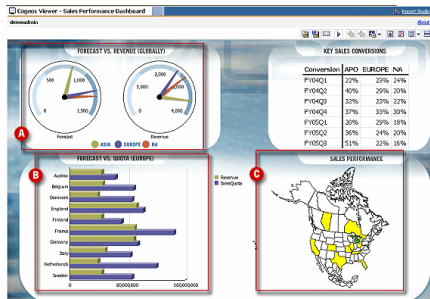
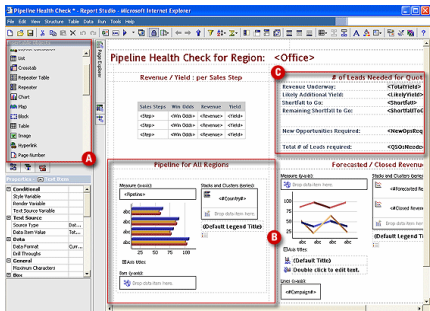
- ▶ Ergänzung um Audio- oder Videodaten
- ▶ Einbeziehung von Dokumentenmanagementsystemen

Analysewerkzeuge: Funktionalität

● Data Access

- ▶ Reporting Werkzeuge
- ▶ Lesen von Daten, Veränderung/Anreicherung durch einfache arithmetische Operationen
- ▶ Präsentation in Berichten
- ▶ „Ampelfunktionen“: regelgebundene Formatierung
- ▶ Basis: SQL

Analysewerkzeuge: Beispiel [Cognos, 2006]



Analysewerkzeuge: Funktionalität

● OLAP

- ▶ Interaktive Datenanalyse, Klassifikationsnavigation
- ▶ Berichte mit verdichteten Werten (Kennzahlen)
- ▶ Navigationsoperationen:
 - ★ Drill Down,
 - ★ Roll Up,
 - ★ Drill Across,
 - ★ Dice und
 - ★ Slice
- ▶ Gruppierungs- und Berechnungsfunktionen (statistisch, betriebswirtschaftlich)
- ▶ Validierung von Hypothesen, Plausibilitätsprüfung

Analysewerkzeuge: Beispiel [Cognos, 2006]

REVENUE BY: ORDER METHOD VERSUS PRODUCT (TOP 3 REGIONS) - Analysis Studio - Microsoft Internet Explorer

Insertable Objects: Go Sales (PC)

- Years
 - Current Month
 - Last Month
 - QTD
 - Prior QTD
 - QTD Grouped
 - YTD
 - Prior YTD
 - YTD Grouped
- Products
- Staff
- Sales Territory
- Retailer
 - By Retailer site (10 of 21)
 - Sweden
 - Japan
 - Taiwan
 - Korea
 - China
 - Australia
 - Belgium
 - Denmark

Rows: Combination Columns: QTD Grouped Context: Products (list)

Filters are applied. See the Properties pane for more details.

Revenue	Prior QTD	QTD	QTD Change	QTD Growth
Telephone	\$2,437,167.60	\$5,472,020.62	\$3,034,853.02	124.52%
E-mail	\$1,394,147.12	\$3,344,524.16	\$1,950,377.04	139.56%
Sales visit	\$3,406,518.14	\$4,863,403.16	\$1,376,970.02	39.49%
Web	\$4,130,883.24	\$5,296,920.22	\$1,166,036.98	28.27%
Fax	\$754,480.44	\$1,280,583.10	\$526,102.66	69.72%
Mail	\$342,898.46	\$463,937.80	\$121,039.34	35.30%
Special	\$243,098.32	\$295,017.00	\$51,918.60	21.36%
Order Method	\$12,789,193.32	\$21,016,491.86	\$8,227,298.54	64.27%
Canada	\$888,852.88	\$2,258,730.16	\$1,369,877.28	154.12%
United States	\$1,193,602.78	\$2,047,787.12	\$854,184.34	71.58%
Brazil	\$288,350.30	\$907,395.94	\$619,045.64	214.69%
Subtotal (included)	\$2,370,805.96	\$5,213,903.22	\$2,843,097.26	119.92%
Subtotal (excluded)	\$4,999,571.18	\$7,672,659.74	\$2,673,088.56	53.47%
By Retailer site	\$7,370,377.14	\$12,886,562.96	\$5,516,185.82	74.84%
Italy	\$58,133.58	\$267,113.70	\$208,980.12	359.49%
United Kingdom	\$65,823.68	\$224,920.54	\$159,104.06	241.71%
Canada	\$195,700.10	\$339,549.76	\$143,849.66	73.56%
Subtotal (included)	\$319,657.36	\$831,692.00	\$512,034.64	160.18%
Subtotal (excluded)	\$1,632,812.88	\$2,369,222.86	\$736,409.98	45.11%
By Retailer site	\$2,152,470.24	\$3,200,914.86	\$1,048,444.62	48.71%
Golf Equipment				

Information - Penelope Tamrine

Name Penelope Tamrine

Level General manager

Aggregation Rollup

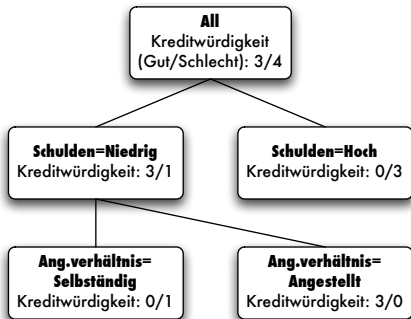
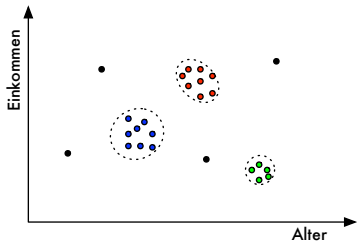
REVENUE BY: ORDER METHOD VERSUS PRODUCT (TOP 3 REGIONS)

Analysewerkzeuge: Funktionalität

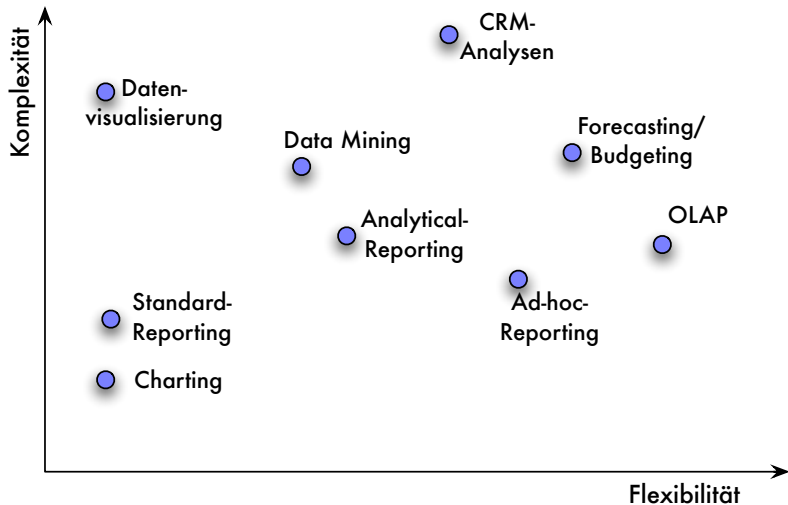
● Data Mining

- ▶ Aufdeckung bisher unbekannter Zusammenhänge
→ Muster, Pfade, Regeln
- ▶ Verfahren (u.a.):
 - ★ Klassifikation: Zuordnung der Daten zu vorgegebenen Klassen
 - ★ Assoziationsregeln
 - ★ Clusterbildung: Segmentierung, d.h. Daten bzgl. Ihrer Merkmalsausprägungen zu Gruppen zusammenfassen
 - ★ Prognose

Data Mining: Beispiele



Arten von Analysen



Analysewerkzeuge: Realisierung

- Standard Reporting:
 - ▶ Reporting-Werkzeuge des klassischen Berichtswesens
- Berichtshefte:
 - ▶ Graphische Entwicklungsumgebungen zur Erstellung von Präsentationen von Tabellen, Graphiken, etc.
- Ad-hoc Query & Reporting:
 - ▶ Werkzeuge zur Erstellung und Präsentation von Berichten
 - ▶ Verbergen von Datenbankbindung und Anfragesprachen

Analysewerkzeuge: Realisierung

- Analyse-Clients:
 - ▶ Werkzeuge zur mehrdimensionalen Analyse
 - ▶ Beinhalten Navigation, Manipulation (Berechnung), erweiterte Analysefunktionen und Präsentation
- Spreadsheet Add-Ins:
 - ▶ Erweiterung von Tabellenkalkulationen für Datenanbindung und Navigation
- Entwicklungsumgebungen:
 - ▶ Unterstützung der Entwicklung eigener Analyseanwendungen
 - ▶ Bereitstellung von Operationen auf multidimensionalen Daten

Repository

- Aufgabe:
 - ▶ Speicherung der Metadaten des DW-Systems
- Metadaten:
 - ▶ Informationen, die Aufbau, Wartung und Administration des DW-Systems vereinfachen und Informationsgewinnung ermöglichen
 - ▶ Beispiele:
 - ★ Datenbankschemata,
 - ★ Zugriffsrechte,
 - ★ Prozessinformationen (Verarbeitungsschritte und Parameter), etc.

Metadaten-Manager

- Aufgaben:
 - ▶ Steuerung der Metadatenverwaltung
 - ▶ Zugriff, Anfrage, Navigation
 - ▶ Versions- und Konfigurationsverwaltung
- Formen:
 - ▶ Allgemein einsetzbar: erweiterbares Basisschema
 - ▶ Werkzeugspezifisch: fester Teil von Werkzeugen
- Häufig Integration von bzw. Austausch zwischen dezentralen Metadaten-Managementsystemen notwendig

Zusammenfassung

- Referenzarchitektur für Data-Warehouse-Systeme
- Prozess des Data Warehousing
- Rolle der Komponenten
- Data Marts als Extrakte des DW
- Analysewerkzeuge: Klassifikation und Beispiele