Merkblatt DWH

Mittwoch, 6. Januar 2016 13:55

Version: 1.0.0
Study: 3. Semester, Bachelor in Business and Computer Science
School: Hochschule Luzern - Wirtschaft
Author: Janik von Rotz (http://janikvonrotz.ch)

A data warehouse (DWH) is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of managements decision making process.

Data Warehousing: Prozesse zur Erstellung, Bestückung, Bewirtschaftung, Verwendung von DWHs.

Operational Data Store (ODS): Ist ein DWH-Typ. Ist eine zwischenstufe zwischen Quellsystemen und dem DWH. Wird für Kurzzeitanalysen genutzt. Definition Skritp S.32.

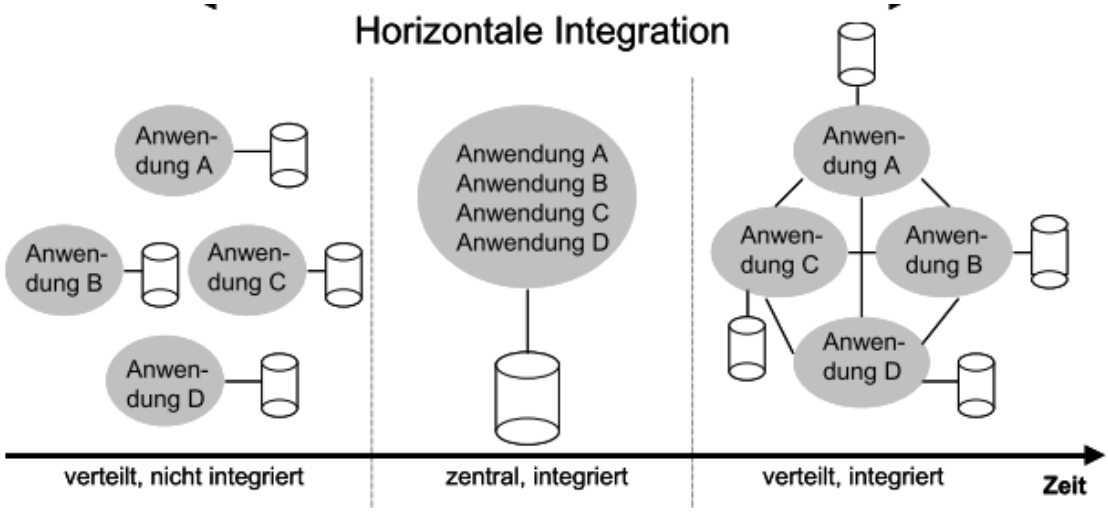
Data Marts

Ein Data Mart (DM) ist eine spzialisierte analytische Datenbank für eine Abteilung, eine Arbeitsgrupp, eine Einzelperson oder für die Daten einer umfangreichen Applikation.

DMs sind
- Einfache Datenmodelle
- Zugriffsoptimiert
- Werden dezentral in Abteilungen gepflegt

Ausführungen
- Bottom up: DWH entsteht durch integration bestehender DM.
- Parallelität: DWH und DM werden zu einem definierten Grad der Unabhängigkeit entwickelt und liefern sich gegenseitig Daten.
- **Top Down**: DM erhalten Daten aus zentralem DWH

Architektur



Datenbeschaffung

ETL Prozess: Daten werden aus Datenquellen extrahiert und in der Staging-Area zwischengespeichert. Dort werden Datenbereinigt (cleanse, scrub) und transformiert (transform). Die aufbereiteten Daten werden in die Zieldatenbank gespeichert (load).

Daten werden in diesem Prozess
- Gefiltert
- Integriert
- Bereinigt

- Zugeordnet
- Homogenisiert
- Angereichert
- Verdichtet
- Aggregiert
- konsolidiert

Extraktion
Lädt Daten aus Quellen in den Arbeitsbereich.
Das erfolgt
- Perodisch
- Auf abruf
- Ereignisgesteuer
- Nach Mutation, sofort

Quellen sind OLTP Systeme.
On-Line Transactional Processing: Operative Datenbanksysteme.

Extraktionsprozess wird von Monitor überwacht

Arbeisbereich (staging area)
Temporäre Datenhaltungskomponente für die Datenaktualisierung in der Basisdatenbank.

Transformation
Daten werden
- Struktuerell
- Semantisch
- Homgenisiert

Das Data Cleansing umfasst:
- Daten nachtragen
- Dubletten eliminieren
- Fehler beseitigen
- Aktualisieren

Herausforderung ist ein sauberes Delta mit Historisierung zu generieren.

Ladekomponente
Übertragung der integrierten, homogenisieren, bereinigten und bereicherten Daten in die Basisdatenbank.

Auswertungsbereich
Ladekomponente hat die Basisdatenbank bestückt. Die Datenablage ist
- Modellmässig unabhängig (von den Quellen)
- Feingranular
- Aktuell
- Für Analysezwecke ausgelegt
- Universell einsetzbar

Metadaten

Definiert und beschreibt die Struktur, Operationen und Inhalt eines Informationssystems.

Werden in Repositorium gespeichert.

Technische Metadaten
- Datenmodell der Quellsysteme

- Data Cleansing Rules
- Erstellungsdatum
- Spaltenname (ID, name)
- Datenbankname
- Beziehungen
- Domänen
- Systeminventur

Business Metadaten

- Minimaler Umsatz -> Geschätsregel
- Data Mart Verkauf
- Kennzahl Umsatz
- Daten Transformationsregeln
- Reporting Tools
- Currency OLAP Data
- Gruppierungen
- Aggregationen

Datenqualität

Daten entsprechen in der Gesamtheit von Eigesnchaften und Merkmalen den Anforderungen an den Datenbestand.

Ursache und Orte von Qualitätsmängeln

- Schlechte Datenerfassung infolge Ignoranz
- Schlechte Prozesse
- Mangelde Architektur
- Unzureichende Definitionen
- Unpassende Datenverwendung
- Datenverfall durch mangelnde Pfelege

Datenqualität ist Faktor Mensch entscheiden.

Es ist ein Frage von

- Sachwissen
- Sorgfalt
- Kommunikation
- Kompetenzen
- Identifikation
- Belobigung, Incetivierung
- Saktionen

Ob die die Daten korrekt erfasst werden.

Qualitätsmangel kann im ganzen ETL und Auswertungsprozess auftreten.

Metrik für Datenqualität

- Kriterien
- Erfüllungsgrad

Data Profiling

Ziel ist des die Daten einer Unternehmung zu kennen.

Data Profiling erfassen von Metadaten zu Datenquellen

- Herkunft
- Struktur
- Quellformat

- Menge
- Benennungen
- Qualität
- Zweck
- Zustand

Methoden des Profilings sind
- Deskriptiv (beschreibend): Analyse von Häufigkeiten, Abhängigkeiten, Aussreissern
- Kognitiv (lernen): Regelinduktion, Klassifizierungen
- Deduktiv (ableitend: Regelanalyse

Regeltypen
- deterministisch: nur ab 12 Jahren
- Stoachaistisch: Wahrscheinlichkeit -> Ab 60 keine Kinder gebären

Redundanz kann auftreten als
- Dubletten -> vollständig identisch
- Ähnlich bis zu einem bestimmten Grad

Mit Distanzmasse bestimmen ob es Duplikat ist oder nicht.

Levenshtein-Distanz: Anzahl Schritte, die nötig sind, um die Zeichenkette in Zeichenkette B zu überführen.

T**ier**, Tor = 2	1 Z verändern, 1 Z löschen

Datawarehousing Porzesse

Vorbereitung
Für die ETL Prozesse ist unverzeichtbar, ddass vorher
- Datenschreibung (Profiling)
- Schemaintegration

Historisierung

SCD Typ 1: Keine Historisierung
SCD Typ 2: Satzweite Speicherung
- Neue Attribute: dat_von, dat_bis, gültig, vorher_id
- Ist dat_bis NULL dann ist dies der aktuelle Datensatz

SCD Typ3
- Nur neue und alte Information wird mithilfe zusätzlicher Spalte gespeichert.

OLAP und ERM

On-Line Analytical Processing: Systeme zur Geschäftsanalyse und Entscheidungsfindung

Fakten sind
- Kennzahlen
- Skalar
- Zahlen

Dimensionen sind
- Deskriptiv
- Elemente
- Vielmals Text

Beispiele für Hierarchische Dimensionen

- Produkt, Produktgruppe, Produktfamilie, Produktportfolio
- Standort, Strasse, Ortschaft, Bezirk, State, Country

MOLAP: Multidimensionale Systeme
ROLAP: Relationale Datenbanksysteme

Operationen am Würfel

- Slicing: Eine Bedingung -> Eine Scheibe des Würfels
- Dicing: Drehung einer Dimensionsachse
- Roll-Up: Bewegung entlang der Elementhierarchie -> Granularität wird vergröbert
- Drill-Down: Verfeinerung der Granulariät
- Drill-Across: Kombination der Cubes

Berechnung Rollup, Cube und Grouping Set

Datensätze

Name	Anzahl
Type	2
Store	9
Number	2

Rollup

Type	Store	Number	Summe
0	0	0	2*9*2=36
0	0	1	2*9=18
0	1	0	2*2=4
0	1	1	2
1	0	0	2*9=18
1	0	1	9
1	1	0	2
1	1	1	1

Summe: 90

Formel: a*b*c+a*b+a*c+b*c+a+b+c+1

Cube

Type	Store	Number	Summe
0	0	0	2*9*2=36
0	0	1	2*9=18
0	1	1	2
1	1	1	1

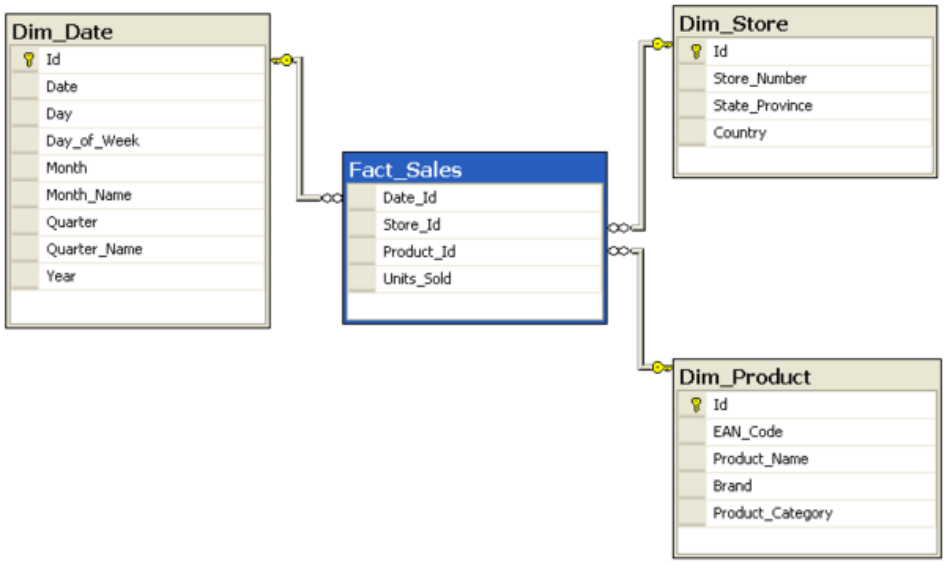
Summe: 57

Grouping Set (Store, Number)
Summe: 2+9=11

Datenmodell Schemas

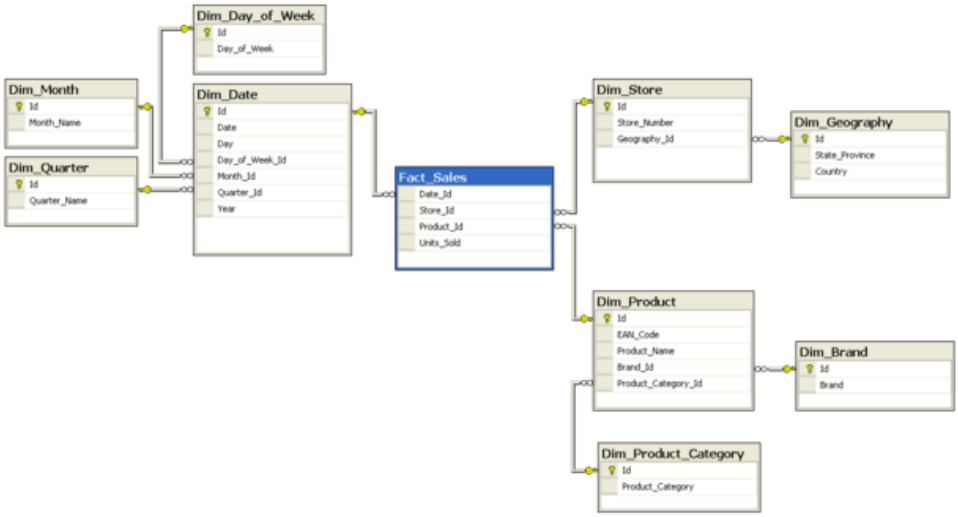
Stern/ Star

- Anfällig für Anomalien
- Performant
- Denormalisiert
- Braucht viel speicher
- Geeignet für Data Marts

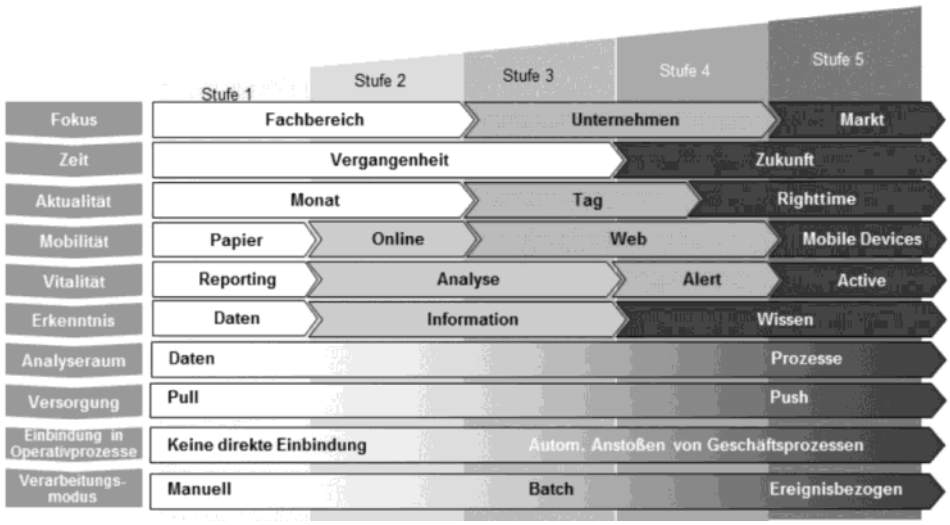


Schneeflocke / Snowflake

- Normalisiert
- Einfacher zur Wartung
- Unperformant -> Viele Joins
- Geeignet für DWH core



Reifegradmodell



Level 0: Limited BI

- Isolierte Informationsbestände
- Beispiel: Excel Spreadsheet

Level 1: Operational Reporting

- Einzelinformation
- Fachabteilungen erzeugen Einzelauswertungen
- Keine Standardisierung
- Beispiel: Access Reports

Level 2: Query & Analysis

- Informationsinseln
- Synergien auf Fachbereichsebene
- Lösungen laufen stabil
- Ad-hoch-Auswertungsfunktionalität
- Beispiel: Erweiterte Analysetools

Level 3: Dashboard Management

- Informationsintegration
- Unternehmensweite Auswertungen
- Standardisierte Lösungen
- Integration der Daten
- Solide Data-Warehouse-Architektur im Einsatz
- Beispiel: Dashboard

Level 4: OLAP

- Information Intelligence
- Unternehmensbereiche arbeiten mit DWH
- Fortgeschrittene Auswertungsverfahren
- Strategische Orientierung
- Beispiel: OLAP, MS Analysis Services

Level 5: Data Mining

- Enterprise Information Management
- Vollständige Integration auswertungsorientierter und operativer Systeme
- Optimale Unterstützung der Geschäftsprozesse
- DWH ist ein unverzichtbares Instrument
- Beispiel: Data Mining, Predictive Analytics

Multiprocessing

Grid and Massiv Parallel Processing/ Clustering

Gemeinsam

- Rechnerverbund
- Koordinierende Instanz
- Ausfalltoleranz

Unterscheide

Grid	MPP
• Dezentral • Plattform Heterogenität • Koordinierendes Betriebssystem	• Zentral • Homogenität • Koordinierender Dienst

Cloud Computing - Outsourcing

Vorteil

- Verfügbarkeit
- Flexible Kosten
- Produktsynergien
- Weniger Verantwortung

Nachteil

- Datenschutz
- Transparenz
- Kontrolle
- Systemabhängigkeit
- Verlust Kompetenzen