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 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.

Data Marts

Ein Data Mart 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

Datenbeschaffung

ETL Prozess

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

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.

Metadaten

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

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

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

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.

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
- Neur neue und alte Information wird mithilfe zusätzlicher Spalte gespeichert.

OLAP und ERM

Fakten sind
- Kennzahlen
- Skalar
- Zahlen

Dimensionen sind
- Deskriptiv
- Elemente
- Vielmals Text

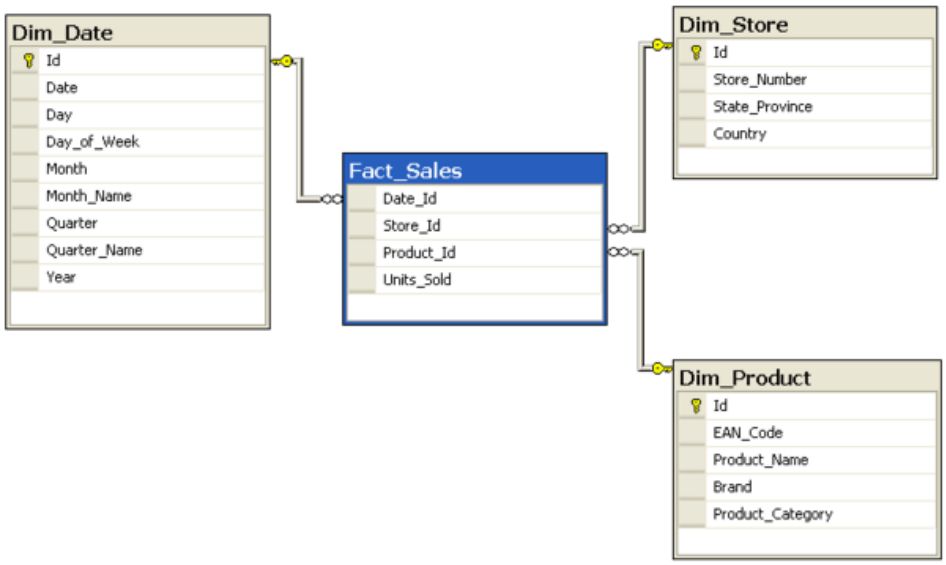
Beispiele für Hierarchische Dimensionen
- Produkt, Produktgruppe, Produktfamilie, Produktportfolio
- Standort, Strasse, Ortschaft, Bezirk, State, Country
-

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

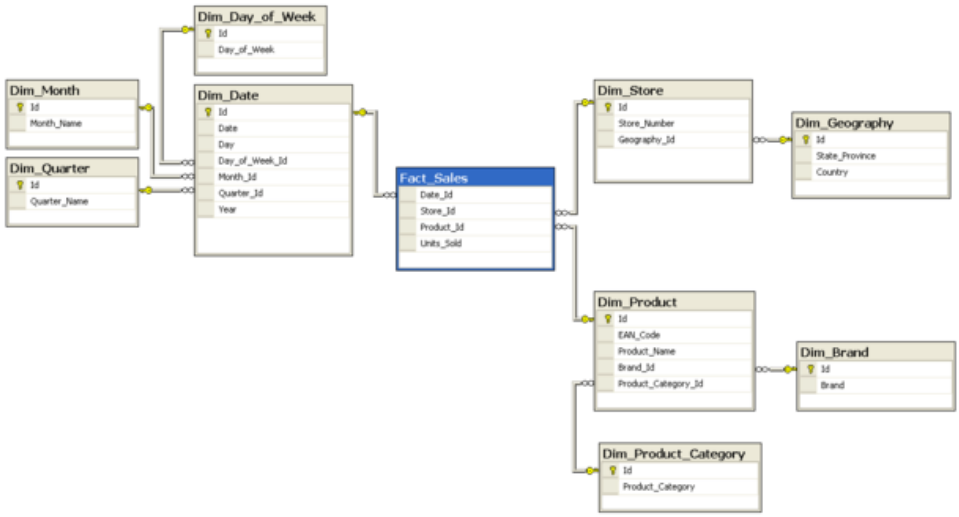
Datenmodell Schemas

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



Multiprocessing

Grid and Massiv Parallel Processing/ Clustering

Gemeinsam

- Rechnerberband
- Koordinierende Instanz
- Ausfalltoleranz

Unterscheide

Grid	MPP
• Dezentral • Plattform Heterogenität	• Zentral • Homogenität