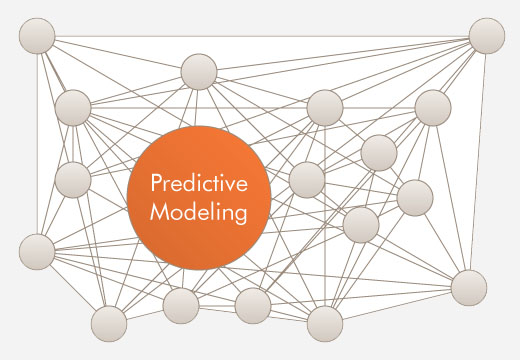
Erklären Sie die folgenden Begriffe (wenn möglich mit grafischer Unterlegung -> Skizze):

* Predective Modelling
* Dependency Modelling
* Data Clustering, (B) Data Summarization, (C) Change Deviation Detection
* Cube
* Cube: (A) Drill Down, (B) Roll Up, (C) Drill Across
* Cube: Slicing / Dicing
* Basic Fact Table - Aggregated Fact Table
* Star Schema, (B) Snowflake Schema, (C) Fact Constellation Schema, (D) Galaxy Schema, (E) UNIMU Schema

Predective Modelling

Auf Deutsch Prädiktive Modellierung, verwendet vorhandene Daten, um zukünftige Ereignisse vorherzusagen. Es werden vorhandene Daten benutz, um ein Modell zu erstellen, das wichtige Trends erfasst. Dieses prädiktive Modell wird dann auf aktuelle Daten angewendet, um vorherzusagen, was als Nächstes passieren wird, oder um Aktion vorzuschlagen, mit denen optimale Ergebnisse erreicht werden können.



Dependency Modelling

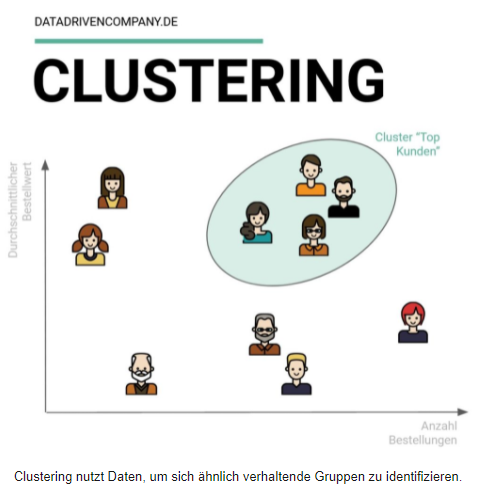
Ein Abhängigkeitsmodell basiert auf Zielen und Zielsetzungen und den Voraussetzungen, diese Ziele zu erfüllen. Es ist ein positivistischer Top-Down-Ansatz, der von Zielen zu Anforderungen arbeitet. Das Prinzip hinter Dependency Modeling ist, dass der Analyst idealerweise, nicht all die Dinge herausfinden muss, die in seinem Unternehmen schiefgehen könnten. Seine Verantwortung sollte sich darauf beschränken, ein ausreichend genaues Modell des Unternehmens bereitzustellen, und die Software erledigt den Rest. Beim Dependency Modeling bauen man also ein abstraktes Modell des Unternehmens auf, das die Software dann analysiert.

Der Aufbau eines Abhängigkeitsmodells bietet eine Vielzahl von Vorteilen, wie zum Beispiel:

* Es bildet eine Sprache, um mit anderen über Risiken zu diskutieren.
* Es zwingt uns zu verstehen und zu artikulieren, was wir erreichen wollen.
* Eventuelle Missverständnisse werden für uns und andere sichtbar gemacht und dadurch eher aufgedeckt.
* Es ermöglicht uns, Risiken zu analysieren.

Data Clustering

Unter Cluster versteht man eine Methode, um Datenobjekte in Gruppe zu ordnen. Dabei werden Ähnlichkeiten der Daten genutzt, um möglichst homogene Gruppen zu identifizieren. Clustering arbeitet hierbei ohne vorhanden Wissen, welche Einträge sich ähnlich sind, sondern berechnet die Ähnlichkeiten rein auf der Datengrundlage. Daher ist Clustering eine geeignete Methode ohne Vorwissen Gruppen oder Segmente zu generieren und daraus Wissen abzuleiten.



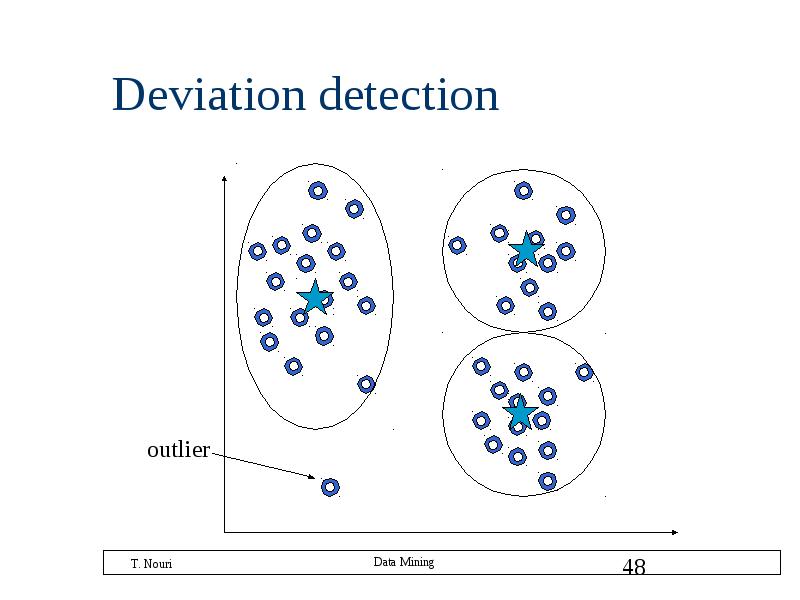
Data Clustering versucht zwei Ziele zu verfolgen: ein Ziel ist es ähnliche Datenpunkte zu kombinieren und somit Komplexität zu verringern. Das andere Ziel ist es, die Datenpunkte zu identifizieren, die eben nicht zu einer großen Gruppe gehören und damit Besonderheiten aufweisen (Outlier-Detection).

Data Summarization

Die Datenzusammenfassung fasst riesige Mengen an Daten zusammen. Da die Daten im Data Warehouse sehr umfangreich sind, muss es einen Mechanismus geben, um nur die relevanten und sinnvollen Informationen in einem einfachen und übersichtlichen Format zu erhalten. Die Datenverdichtung bietet die Möglichkeit, den Benutzer eine allgemeine Sicht auf disparate Datenmengen zu geben. Die Zusammenfassung von Daten in sehr großen mehrdimensionalen Datenbeständen, wie sie in Data Warehouses vorkommen, ist eine sehr anspruchsvolle Aufgabe. Dies erfordert in der Regel sehr intensive Untersuchungen durch IT-Experten, Datenbankadministratoren und Programmierer, damit allgemeine Trends und wichtige Ausnahmen ermittelt und technisch bearbeitet werden können.

Change Deviation Detection

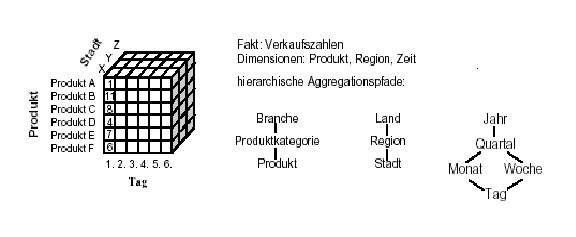
Erkennungen von Abweichungen in Daten ist in den Datenbanken von Bedeutung. Abweichungen werden oft als Ausreißer, Fehler oder Rauschen in den Daten betrachtet. Ziel ist es, ungewöhnliche oder verdächtige Fälle aufgrund von Abweichungen von der Norm in Daten zu erkennen. Das Modell beobachtet Daten, die auf irgendeiner Weise "abnormal" oder "verdächtig" ist.



Cube

Zur Erleichterung der komplexen Analysen und zur Visualisierung der Datenbestände in OLAP-Systemen, werden diese meist als multidimensionale Würfel (data cube) modelliert und dargestellt. Im Mittelpunkt des multidimensionales Modell Ansatzes stehen eine Menge von numerischen Maßzahlen, die sogenannten Fakten, wie z.B. Verkäufe, ROI (Return of Investment), Bestände, Forderungen (vgl. Merkmal Themenorientierung). Jede dieser Maßzahlen ist von einer Menge von Dimensionen abhängig, welche ihren Kontext repräsentiert. Unter anderem können z.B. Produktname, Stadt und Tag des Verkaufs als Dimensionen der Maßzahl Verkäufe angesehen werden. Die Dimensionen determinieren zusammen eindeutig die zugehörige Maßzahl. Jede Dimension wird dabei durch eine Menge von Attributen beschrieben. Die Produkt-Dimension könnte z.B. aus den folgenden drei Attributen bestehen: Kategorie und Branche des Produktes, sowie das Jahr seiner Einführung. Diese Attribute können innerhalb einer Dimension miteinander in Beziehung stehen und so eine Beziehungshierarchie bilden.

Es stehen also der Produktname mit der Produktkategorie und diese wiederum mit der Branche in solch einer hierarchischen Beziehung.



Auf einem multidimensionalen Würfel sind neben den normalen Datenbankoperationen noch weitere, für OLAP spezifische Operationen definiert:

* Slice-and-Dice:   
  Bestimmte Ausschnitte der Daten können in jeder beliebigen Dimension geschnitten oder gedreht werden. Im Beispiel können z.B. nur die Verkäufe von Produkt A für verschiedene Tage und Städte betrachtet werden. Dies entspricht einem Schnitt durch den multidimensionalen Würfel.

Ein Bild, das Text, Elektronik enthält.

Automatisch generierte Beschreibung

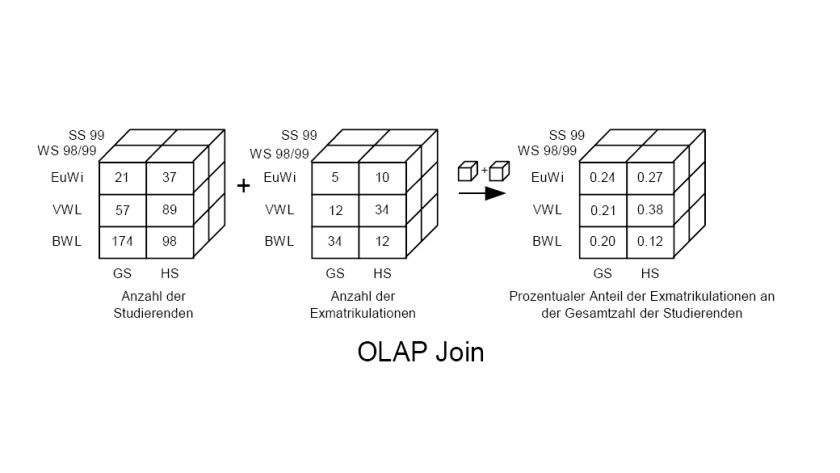
* Drill-Down:   
  Diese Operation erlaubt ein Navigieren entlang der Hierarchien, die für die einzelnen Dimensionen definiert sind. Der Nutzer, der z.B. gerade die Verkäufe auf Landesebene analysiert, kann mittels Drill-Down, die Daten auf Regionen-Ebene herunterbrechen und die Daten so weiter auffächern.
* Roll-Up:   
  Mit dieser Operation kann man das gleiche wie Drill-Down machen, aber nur umgekehrt. Hier werden alle Einzelwerte zu einem weiter oben liegenden Hierarchieattribut verdichtet. Dadurch verringert sich der Detaillierungsgrad.

Ein Bild, das Text, Elektronik enthält.

Automatisch generierte Beschreibung

* Drill-Across:   
  Drill Through bedeutet, dass man durch denselben Bericht drillt, um die Daten zu erhalten.

Drill Across bedeutet, wenn man von einem Bericht zum anderen springt oder navigiert, dann drillt man in den Berichten. Dies erhaltet man über die Eigenschaftsnavigation, und dadurch kann man dann die Daten aus anderen Berichten anzeigen.



Basic Fact Table - Aggregated Fact Table

Im Data Warehousing besteht eine Faktentabelle aus den Messungen, Metriken oder Fakten eines Geschäftsprozesses. Sie befindet sich in der Mitte eines Stern- oder Schneeflockenschemas, das von Dimensionstabellen umgeben ist. Wenn mehrere Faktentabellen verwendet werden, werden diese in einem Faktenkonstellationsschema angeordnet. Eine Faktentabelle hat in der Regel zwei Arten von Spalten: Spalten, die Fakten enthalten, und Spalten, die einen Fremdschlüssel zu Dimensionstabellen darstellen. Der Primärschlüssel einer Faktentabelle ist in der Regel ein zusammengesetzter Schlüssel, der sich aus allen seinen Fremdschlüsseln zusammensetzt. Faktentabellen enthalten den Inhalt des Data Warehouse und speichern verscahiedene Arten von Kennzahlen wie additive, nicht-additive und semiadditive Kennzahlen.

Aggregate sind die Zusammenfassung faktenbezogener Daten zum Zweck einer besseren Leistung. Es gibt viele Fälle, in denen der Kunde schnell eine Frage beantworten möchte, bei der die Daten auf einer höheren Ebene liegen. Um langsame Antworten zu vermeiden, weil das System Daten auf höheren Ebenen zusammenfasst, kann man diese Anfragen vorberechnen und die zusammengefassten Daten in einem separaten Stern speichern. Aggregate können als konforme Faktentabellen betrachtet werden, da sie die gleichen Abfrageergebnisse liefern müssen wie die detaillierte Faktentabelle.

Star Schema

Dies ist das einfachste und effektivste Schema in einem Data Warehouse. Eine Faktentabelle in der Mitte, die von mehreren Dimensionstabellen umgeben ist, ähnelt einem Stern.

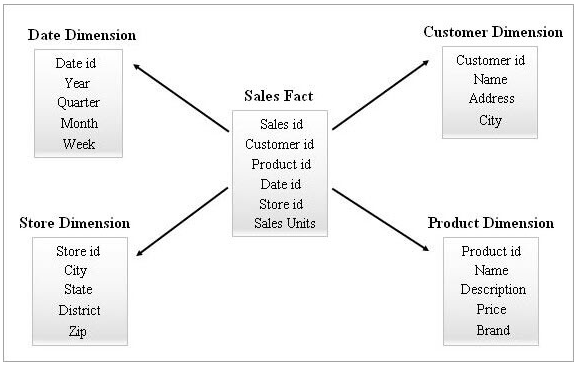
Beim Entwurf von Sternschemata werden die Dimensionstabellen absichtlich de-normalisiert. Sie sind umfangreich mit vielen Attributen, um die kontextbezogenen Daten für eine bessere Analyse und Berichterstattung zu speichern.

Vorteile von Sternschemata

* Abfragen verwenden sehr einfache Joins beim Abrufen der Daten, wodurch die Abfrageleistung erhöht wird.
* Es ist einfach, Daten für die Berichterstattung abzurufen, und zwar zu jedem Zeitpunkt und für jeden Zeitraum.

Nachteile des Star-Schemas

* Wenn sich die Anforderungen häufig ändern, ist es nicht empfehlenswert, das bestehende Sternschema zu ändern und langfristig wiederzuverwenden.
* Die Datenredundanz ist größer, da die Tabellen nicht hierarchisch unterteilt sind.



Snowflake Schema

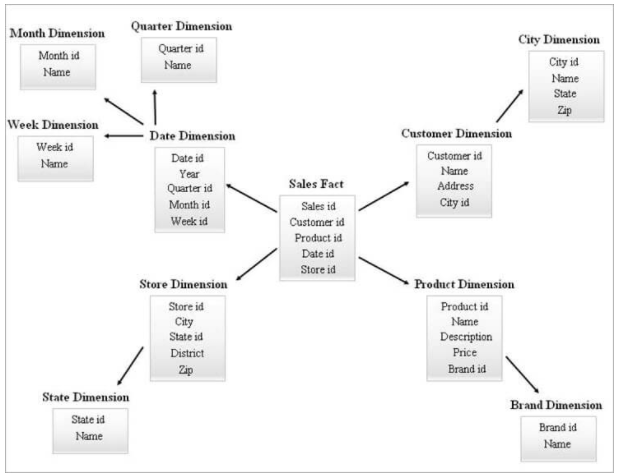
Beim Entwurf von SnowFlake-Schemata werden die Dimensionstabellen absichtlich normalisiert. Zu jeder Ebene der Dimensionstabellen werden Fremdschlüssel hinzugefügt, um eine Verbindung zu ihrem übergeordneten Attribut herzustellen. Die Komplexität des SnowFlake-Schemas ist direkt proportional zu den Hierarchieebenen der Dimensionstabellen.

Vorteile von Sternschemata

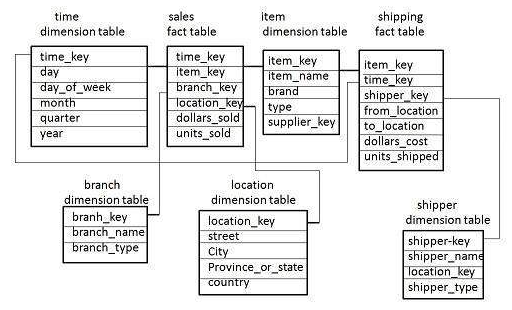
* Die Datenredundanz wird durch die Erstellung neuer Dimensionstabellen vollständig beseitigt.
* Im Vergleich zu einem Sternschema wird durch die Snow Flaking-Dimensionstabellen weniger Speicherplatz benötigt.
* Die Snow Flaking-Tabellen lassen sich leicht aktualisieren (oder pflegen).

Nachteile des Star-Schemas

* Aufgrund der normalisierten Dimensionstabellen muss das ETL-System die Anzahl der Tabellen laden.
* Aufgrund der Anzahl der hinzugefügten Tabellen benötigen Sie möglicherweise komplexe Verknüpfungen, um eine Abfrage durchzuführen. Daher wird die Abfrageleistung beeinträchtigt.



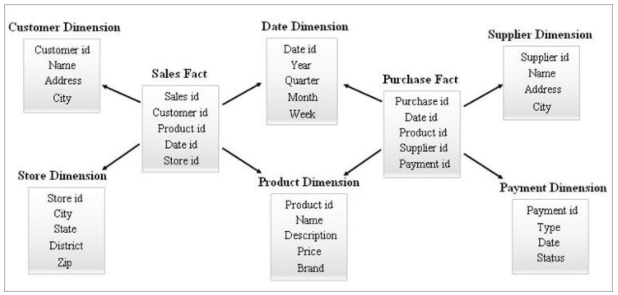
Fact Constellation Schema

Die Faktenkonstellation ist ein Schema zur Darstellung eines mehrdimensionalen Modells. Es ist eine Sammlung von mehreren Faktentabellen mit einigen gemeinsamen Dimensionstabellen. Es ist eines der weit verbreitetsten Schemen für die Entwicklung von Data Warehouses und ist wesentlich komplexer als Stern- und Schneeflockenschemata. Für komplexe Systeme benötigen wir Faktenkonstellationen.

Galxy Schema

diesem Schema teilen sich mehrere Faktentabellen die gleichen Dimensionstabellen. Die Anordnung der Faktentabellen und Dimensionstabellen sieht im Modell des Galaxieschemas wie eine Ansammlung von Sternen aus.

Diese Art von Schema wird für anspruchsvolle Anforderungen und für aggregierte Faktentabellen verwendet, die komplexer sind als das Star-Schema (oder SnowFlake-Schema). Dieses Schema ist aufgrund seiner Komplexität schwer zu pflegen.



UNIMU Schema

Unimu steht für Uniformes Datenschema für multidimensionale Daten und der große Vorteil ist, dass Dimensionen (des Würfel) als Daten und nicht als Datenstrukturen verwaltet werden. Dadurch können neue Kennzahlen, neue Objektdimensionen und neue zeitliche Granularitäten ohne Schemaänderungen abgebildet werden.