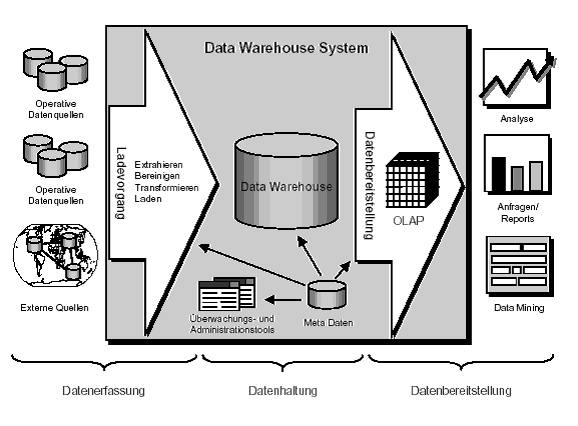
AA6: idealtypische Infrastruktur DWH

Erläutern und beschreiben sie die idealtypische Infrastruktur eines DWH.

Beschreiben Sie insbesondere die folgenden Mechanismen:  
- Datenextraktion aus ERP-Systemen / Datenpumpe  
- Transformation und Bereinigung (OLAP, Data Mining)  
- Aufbau und Funktion der Cubes  
- Slicing, Dicing, Drill-Down



**Architektur von Data Warenhouse Systemen**

Die Architektur lässt sich 3 Ebene unterteilen: Datenerfassung, Datenhaltung und Datenbereitstellung.

**Datenerfassungsebene**

Die Datenerfassungsebene beschäftigt sich mit allen Aufgaben, die mit dem Laden der Daten in das Data Warehouse zusammenhängen. Sie besteht aus einer Vielzahl von Werkzeugen zur Extraktion, Bereinigung und Transformation der Daten, sowie Werkzeugen zum Laden der Daten in das Data Warehouse. Dieser Prozess stellt die Haupteinflussfaktor für eine erfolgreiche Implementierung Eines Data Warehouse Systems dar.

Zum Problembereich der Aktualisierung gehören die Fragen, wann und wie das Data Warehouse zu aktualisieren ist, d.h. wann und wie der Prozess aus Extraktion, Bereinigung / Transformation und Laden anzustoßen ist.

* **Datenextraktion aus ERP-Systemen / Datenpumpe**

Extraktionswerkzeuge sind die Schnittstelle zu den operativen und externen Datenquellen. Sie müssen in der Lage sein, Daten aus den unterschiedlichsten Quellsystemen zu extrahieren, egal ob sie in relationalen, hierarchischen oder Netzwerk-Datenbanken. Dabei sollte sowohl eine komplette Neuberechnung, als auch eine Aktualisierung des Data Warehouses möglich sein.

* Neuberechnung werden sämtliche Daten des Data Warehouses gelöscht und anschließend von Grund auf neu berechnet.
* Die Aktualisierung nutzt hingegen Informationen über die Änderungen auf den Quellsystemen, um die Änderungen für die Daten im Data Warehouse zu berechnen. Es erfolgt also nur eine Aktualisierung der Daten, die sich tatsächlich geändert haben. Aus Performancegründen ist die inkrementelle Aktualisierung eindeutig zu favorisieren.
* **Transformation und Bereinigung (OLAP, Data Mining)**

Da ein Data Warehouse System zur Entscheidungsunterstützung des Managements eingesetzt wird, sind die Anforderungen an die Datenqualität besonders hoch. Die Auswirkungen fehlerhafter Informationen zeigen sich dabei typischerweise sowohl auf der operativen (z.B. Kundenunzufriedenheit), der taktischen (z.B. lange und schlechte Entscheidungsprozesse) und der strategischen Ebene (z.B. Schwierigkeiten bei der Festlegung und Durchsetzung von Strategien). Da Data Warehouse große Datenmengen aus verschiedenen Quellen integriert, ist die Wahrscheinlichkeit für Fehler sehr hoch. Diese Fehler zu entdecken und zu beseitigen, ist die Aufgabe der Extraktions- und Transformationswerkzeuge.

Dabei können grob drei Klassen von Bereinigungstools unterschieden werden:

* Data Migration Werkzeuge führen Transformationen auf den Daten durch. Ein Beispiel hierfür ist z.B. das Ersetzen der Zeichenkette "männlich" durch "m", um ein einheitliches Format für die Speicherung des Geschlechtes eines Kunden im Data Warehouse zu erreichen.
* Data Scrubbing Werkzeuge überprüfen die Zussamenhänge der Daten mit Hilfe domänenspezifischen Wissens (z.B. Postleitzahlenverzeichnisse). Mit Hilfe dieser Tools kann z.B. kontrolliert werden, ob die Stadt und die Postleitzahl in der Adresse eines Kunden zusammenpassen oder ob hier ein Fehler vorliegt.
* Data Auditing Werkzeuge erlauben es Beziehungen und Regelmäßigkeiten in den untersuchten Daten zu entdecken. Auch ist es möglich, Abweichungen von festgelegten Regeln aufzuzeigen. Aus diesem Grund wird diese Klasse oft als Variante von Data Mining Tools bezeichnet.

**Datenhaltungsebene**

Die Aufgabe der Datenhaltungsebene ist die Speicherung der Daten im Data Warehouse. Das Data Warehouse muss in diesem Zusammenhang streng vom OLAP-Speicher unterschieden werden, der die Daten aus dem Data Warehouse übernimmt und sie Analyse- und Reportingwerkzeugen zur Verfügung stellt. Während das Data Warehouse meist relational realisiert wird, sind die Speicherstrukturen des OLAP-Speichers abhängig von der verwendeten Technologie. Man unterscheidet hier zwischen relationalem OLAP (ROLAP), multidimensionalem OLAP (MOLAP) und hybridem OLAP (HOLAP).

Die Datenhaltung in Data Warehouses muss, neben zahlreichen anderen, vor allem zwei Hauptanforderungen genügen. Erstens muss das Data Warehouse in der Lage sein sehr große Datenmengen zu speichern. Zweitens muss es so organisiert sein, dass schnelle Antwortzeiten auf OLAP-Anfragen möglich sind (Die Struktur des Data Warehouses hat Einfluss auf die Struktur des OLAP-Speichers und somit indirekt auf die Anfrageperformance.).

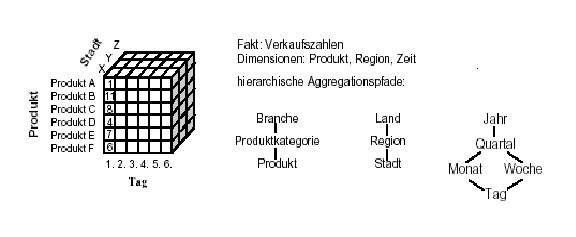
**Datenbereitstellungsebene**

Die Datenbereitstellungsebene besteht aus einem OLAP-Server, der die Daten des OLAP-Speichers an Front End Tools, wie Analyse-, Anfrage/Report-, und Data Mining- Werkzeuge weiterleitet. Auf diesen Daten muss eine multidimensionale Analyse möglich sein.

* **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.
* 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.
* Drill-Up: Mit dieser Operation kann man das gleiche wie Drill-Down machen, aber nur umgekehrt.
* Drill-Through: Mit dieser Operation kann man den zu betrachtenden Datensatz innerhalb einer Hierarchieebene ändern. Der Anwender, der gerade die Daten der Region Nord betrachtet, kann z.B. auf die der Region Süd wechseln.