Business Intelligence

Data Warehouse Concepts

Based on slides by Alejandro Vaisman & Esteban Zimány

Business Intelligence

Contents

- The multidimensional model
- OLAP operations
- Exercises

Imagem retirada de www.demandsolutions.com a 11/03/2015

The Multidimensional Model

OLAP vs OLTP

- Traditional database systems were designed and tuned to support the day-to-day operation:
  - Ensure fast, concurrent access to data
  - Transaction processing and concurrency control
  - Focus on online update data consistency
  - Known as operational databases or online transaction processing (OLTP)

- OLTP DB data characteristics:
  - Detailed data
  - Do not include historical data
  - Highly normalized
  - Poor performance on complex queries including joins an aggregation

- Data analysis requires a new paradigm: online analytical processing (OLAP)
  - Typical OLTP query: pending orders for customer c1
  - Typical OLAP query: total sales amount by product and by customer

OLAP vs OLTP

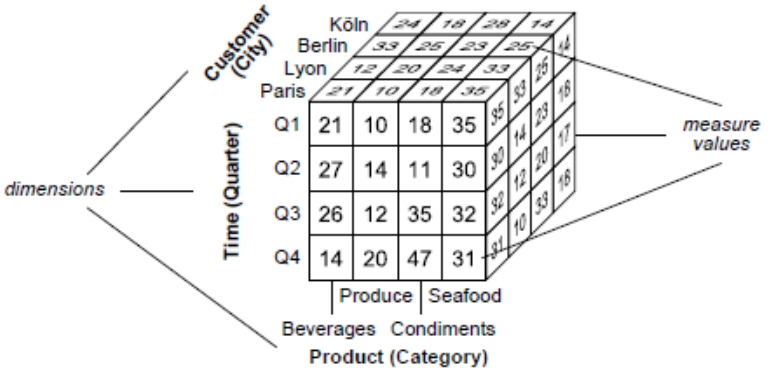
- OLAP characteristics
  - OLTP paradigm focused on transactions, OLAP focused on analytical queries
  - Normalization not good for analytical queries, reconstructing data requires a high number of joins
  - OLAP databases support a heavy query load
  - OLTP indexing techniques not efficient in OLAP: oriented to access few records
    - OLAP queries typically include aggregation

- The need for a different database model to support OLAP was clear: led to data warehouses

- Data warehouse: (usually) large repositories that consolidate data from different sources (internal and external to the organization), are updated online, follow the multidimensional data model, designed and optimized to efficiently support OLAP queries

The Multidimensional Model

- Views data in an n-dimensional space: A data cube
- A data cube is composed of dimensions and facts
- Dimensions: Perspectives used to analyse the data
  - Example: A three-dimensional cube for sales data with dimensions Product, Time, and Customer, and a measure Quantity
- Attributes describe dimensions
  - Product dimension may have attributes ProductNumber and UnitPrice (not shown in the figure)
- The cells or facts of a data cube have associated numeric values called measures
- Each cell of the data cube represents Quantity of units sold by category, quarter, and customer's city

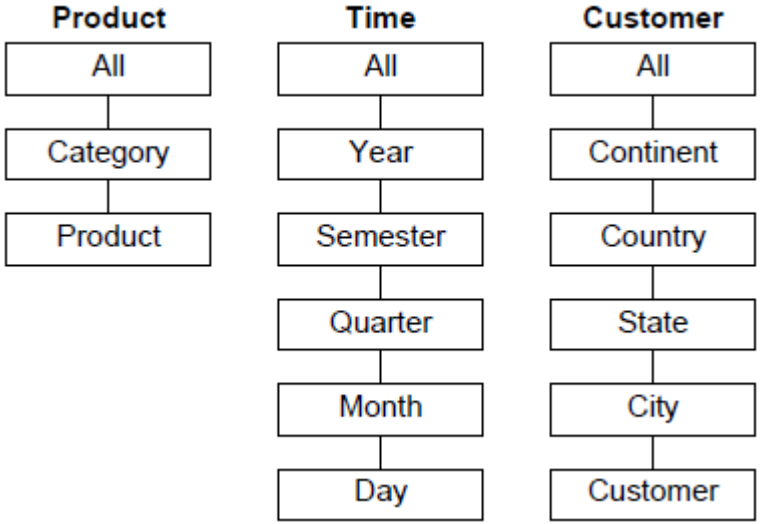


The Multidimensional Model

- Data granularity: level of detail at which measures are represented for each dimension of the cube
  - Example: sales figures aggregated to granularities Category, Quarter, and City
- Instances of a dimension are called members
  - Example: Seafood and Beverages are members of the Product at the granularity Category
- A data cube contains several measures, e.g. amount, indicating the total sales amount (not shown)
- A data cube may be sparse (typical case) or dense
  - Example: not all customers may have ordered products of all categories during all quarters
- Hierarchies: allow viewing data at several granularities
  - Define a sequence of mappings relating lower-level, detailed concepts to higher-level ones
  - The lower level is called the child and the higher level is called the parent
  - The hierarchical structure of a dimension is called the dimension schema
  - A dimension instance comprises all members at all levels in a dimension
  - In the previous figure, granularity of each dimension indicated between parentheses: Category for the Product dimension, Quarter for Time, and City for Customer
  - We may want sales figures at a finer granularity (Month), or at a coarser granularity (Country)

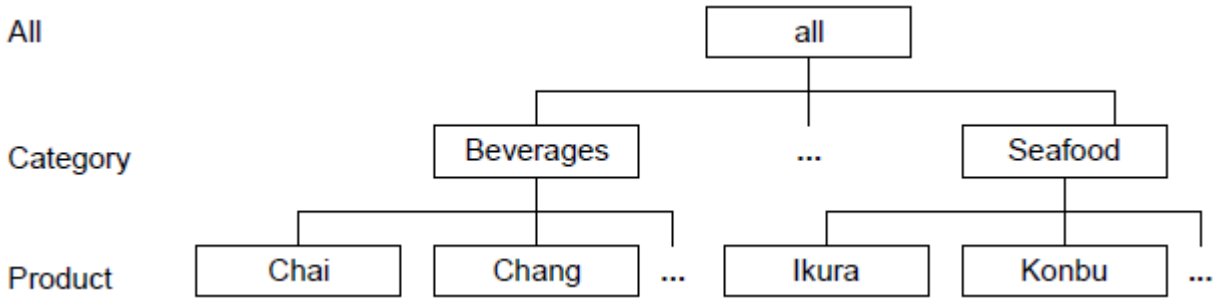
The Multidimensional Model

- Hierarchies of the Product, Time, and Customer dimensions



The Multidimensional Model

- Members of a hierarchy Product → Category



The Multidimensional Model

- Aggregation of measures changes the abstraction level at which data in a cube are visualized

- Measures can be:
  - Additive: can be meaningfully summarized along all the dimensions, using addition
    - The most common type of measures
  - Semi-additive: can be meaningfully summarized using addition along some dimensions
    - Example: inventory quantities, which cannot be added along the Time dimension
  - Non-additive measures cannot be meaningfully summarized using addition across any dimension
    - Example: item price, cost per unit, and exchange rate

- Another classification of measures:
  - Distributive: defined by an aggregation function that can be computed in a distributed way
    - Functions count, sum, minimum, and maximum are distributive, distinct count is not
    - Example: S = {3, 3, 4, 5, 8, 4, 7, 3, 8} partitioned in subsets {3, 3, 4}, {5, 8, 4}, {7, 3, 8} gives a result of 8, while the answer over the original set is 5
  - Algebraic measures are defined by an aggregation function that can be expressed as a scalar function of distributive ones; example: average, computed by dividing the sum by the count
  - Holistic measures cannot be computed from other sub-aggregates (e.g., median, rank)

The Multidimensional Model

- When defining a measure we must determine the associated aggregation functions
  - For example, a semi-additive measure representing inventory quantities can be aggregated using average along the Time dimension, and using addition along other dimensions
- Summarizability refers to the correct aggregation of cube measures along dimension hierarchies
- Summarizability conditions:
  - Disjointness of instances: the grouping of instances in a level with respect to their parent in the next level must result in disjoint subsets
  - Completeness: all instances are included in the hierarchy and each instance is related to one parent in the next level
  - Correctness: refers to the correct use of the aggregation functions (more on this next)