Principles of Knowledge Discovery in Data

Fall 2004

Chapter 2: Data Warehousing and OLAP

Dr. Osmar R. Zaïane

University of Alberta

Summary of Last Chapter

- What kind of information are we collecting?
- What are Data Mining and Knowledge Discovery?
- What kind of data can be mined?
- What can be discovered?
- Is all that is discovered interesting and useful?
- How do we categorize data mining systems?
- What are the issues in Data Mining?
- Are there application examples?

Course Content

- Introduction to Data Mining
- **Data warehousing and OLAP**
- Data cleaning
- Data mining operations
- Data summarization
- Association analysis
- Classification and prediction
- Clustering
- Web Mining
- Similarity Search
- *Other topics if time permits*

Chapter 2 Objectives

Realize the purpose of data warehousing.

Comprehend the data structures behind data warehouses and understand the OLAP technology.

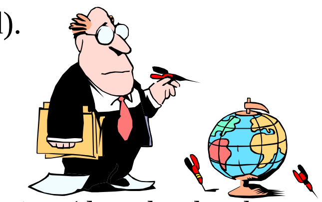
Get an overview of the schemas used for multi-dimensional data.

Data Warehouse and OLAP Outline

- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?

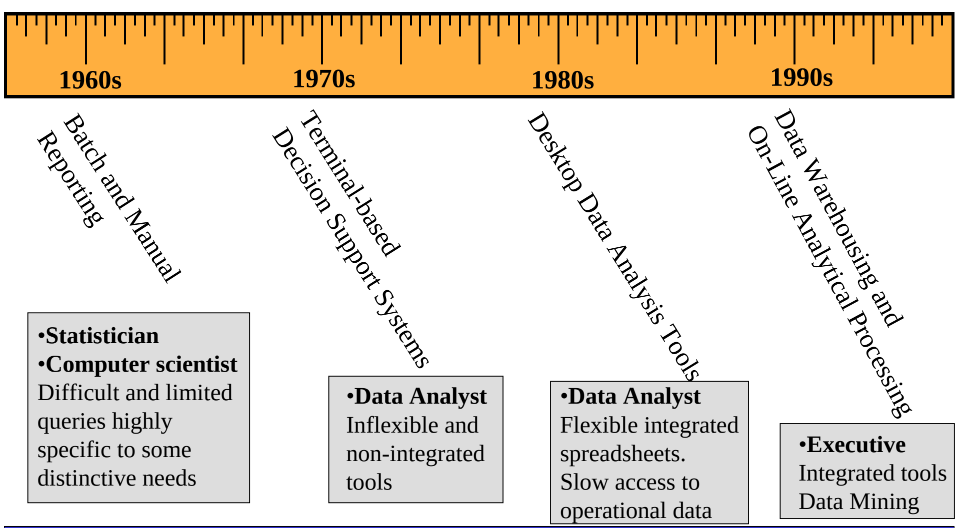
Incentive for a Data Warehouse

- Businesses have a lot of data, operational data and facts.
- This data is usually in different databases and in different physical places.
- Data is available (or archived), but in different formats and locations. (heterogeneous and distributed).



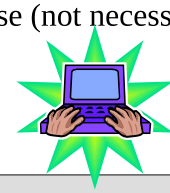
- Decision makers need to access information (data that has been summarized) virtually on one single site.
- This access needs to be fast regardless of the size of the data, and how old the data is.

Evolution of Decision Support Systems

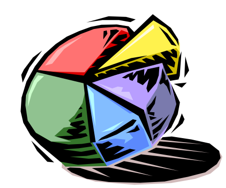


What Is Data Warehouse?

- A data warehouse *consolidates* different data sources.
- A data warehouse is a database that is *different and maintained separately* from an operational database.
- A data warehouse combines and merges information in a consistent database (not necessarily up-to-date) to help decision support.



Decision support systems access data warehouse and do not need to access operational databases → do not unnecessarily over-load operational databases.



Definitions

Data Warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. *(W.H. Inmon)*

Subject oriented: oriented to the major subject areas of the corporation that have been defined in the data model.

Integrated: data collected in a data warehouse originates from different heterogeneous data sources.

Time-variant: The dimension "time" is all-pervading in a data warehouse. The data stored is not the current value, but an evolution of the value in time.

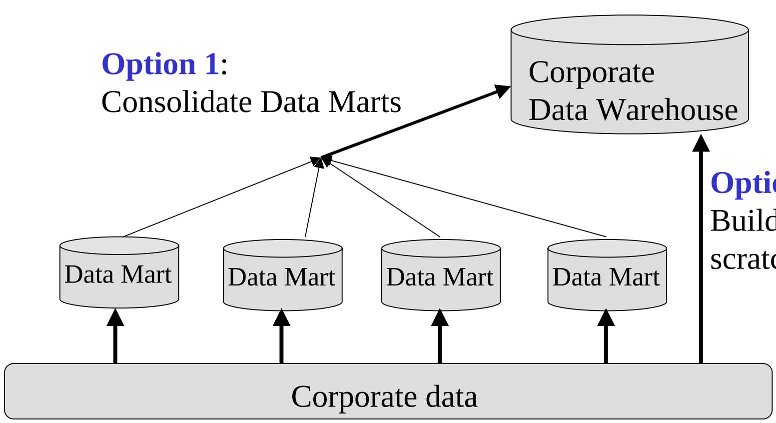
Non-volatile: update of data does not occur frequently in the data warehouse. The data is loaded and accessed.

Definitions (con't)

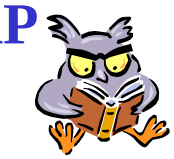
Data Warehousing is the process of constructing and using data warehouses.

A corporate data warehouse collects data about ***subjects*** spanning the **whole** organization. **Data Marts** are specialized, single-line of business warehouses. They collect data for a department or a specific group of people.

Building a Data Warehouse



Data Warehouse and OLAP Outline



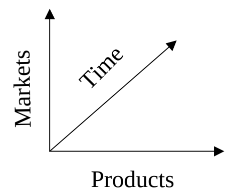
- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?

Describing the Organization

We sell products in various markets, and we measure our performance over time

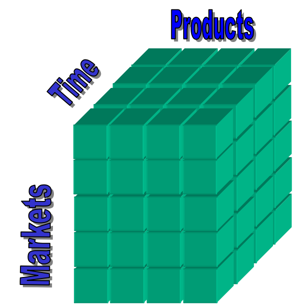
Business Manager

We sell ***Products*** in various ***Markets***, and we measure our performance over ***Time***



Data Warehouse Designer

Construction of Data Warehouse Based on Multi-dimensional Model

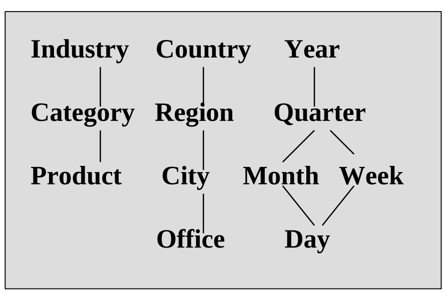


• Think of it as a *cube* with labels on each edge of the cube.
• The cube doesn't just have 3 dimensions, but may have many dimensions (N).
• Any point inside the cube is at the intersection of the coordinates defined by the edge of the cube.
• A point in the cube may store values (measurements) relative to the combination of the labeled dimensions.

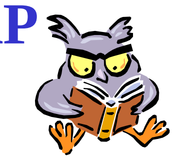
Concept-Hierarchies

Dimensions are hierarchical by nature: total orders or partial orders
Example: Location(continent ➔ country ➔ province ➔ city)
Time(year➔quarter➔(month,week)➔day)

Dimensions: Product, Region, week
Hierarchical summarization paths



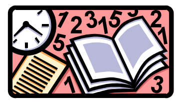
Data Warehouse and OLAP Outline



- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?

On-Line Transaction Processing

• Database management systems are typically used for on-line transaction processing (OLTP)
• OLTP applications normally automate clerical data processing tasks of an organization, like data entry and enquiry, transaction handling, etc. (access, read, update)
• Database is current, and consistency and recoverability are critical. Records are accessed one at a time.



➢OLTP operations are structured and repetitive
➢OLTP operations require detailed and up-to-date data
➢OLTP operations are short, atomic and isolated transactions

Databases tend to be hundreds of Mb to Gb.

On-Line Analytical Processing

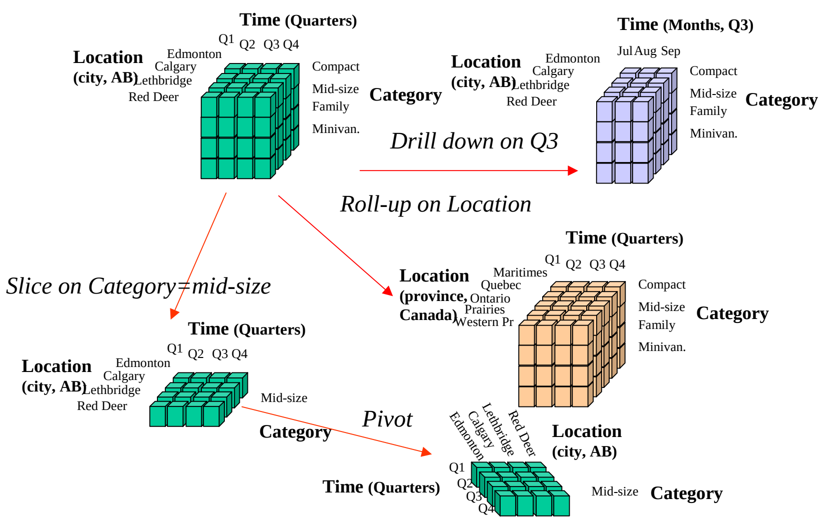
• On-line analytical processing (OLAP) is essential for decision support.
• OLAP is supported by data warehouses.
• Data warehouse consolidation of operational databases.
• The key structure of the data warehouse always contains some element of time.
•Owing to the hierarchical nature of the dimensions, OLAP operations view the data flexibly from different perspectives (different levels of abstractions).

•OLAP operations:

• **roll-up** (increase the level of abstraction)
• **drill-down** (decrease the level of abstraction)
• **slice** and **dice** (selection and projection)
• **pivot** (re-orient the multi-dimensional view)
• **drill-through** (links to the raw data)

DW tend to be in the order of Tb

Data Warehouse OLAP Example



OLTP vs OLAP

	OLTP	OLAP
users	Clerk, IT professional	Knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

(Source: JH)

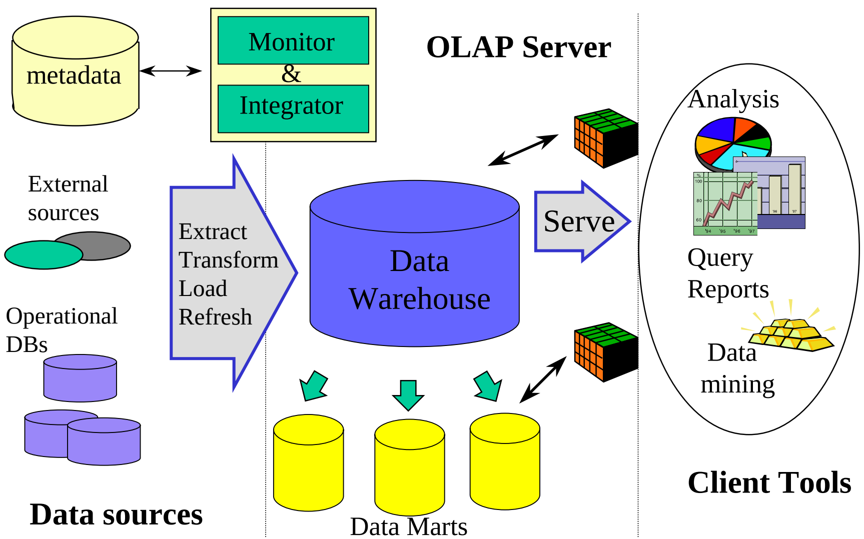
Why Do We Separate DW From DB?

- Performance reasons:
  - OLAP necessitates special data organization that supports multidimensional views.
  - OLAP queries would degrade operational DB.
  - OLAP is read only.
  - No concurrency control and recovery.
- Decision support requires historical data.
- Decision support requires consolidated data.

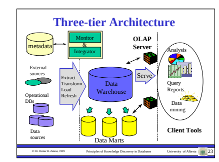
Data Warehouse and OLAP Outline

- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?

Three-tier Architecture

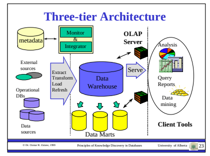


Data Sources



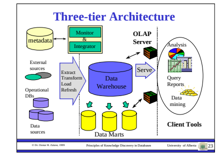
- Data sources are often the operational systems, providing the lowest level of data.
- Data sources are designed for operational use, not for decision support, and the data reflect this fact.
- Multiple data sources are often from different systems run on a wide range of hardware and much of the software is built in-house or highly customized.
- Multiple data sources introduce a large number of issues -- semantic conflicts.

Data Cleaning



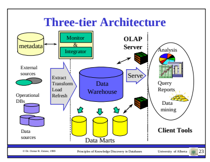
- Data cleaning is important to warehouse.
  - Operational data from multiple sources are often noisy (may contain data that is unnecessary for DS).
- Three classes of tools.
  - Data migration: allows simple data transformation.
  - Data scrubbing: uses domain-specific knowledge to scrub data.
  - Data auditing: discovers rules and relationships by scanning data (detect outliers).

Load and Refresh



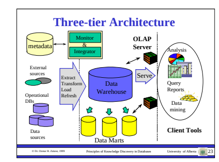
- Loading the warehouse includes some other processing tasks:
  - Checking integrity constraints, sorting, summarizing, build indices, etc.
- Refreshing a warehouse means propagating updates on source data to the data stored in the warehouse.
  - When to refresh.
    - Determined by usage, types of data source, etc.
  - How to refresh.
    - Data shipping: using triggers to update snapshot log table and propagate the updated data to the warehouse.
    - Transaction shipping: shipping the updates in the transaction log.

Monitor



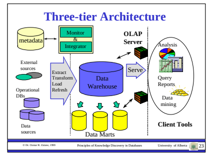
- Detect changes to an information source that are of interest to the warehouse.
  - Define triggers in a full-functionality DBMS.
  - Examine the updates in the log file.
  - Write programs for legacy systems.
- Propagate the change in a generic form to the *integrator*.

Integrator



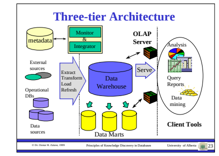
- Receive changes from the monitors
  - Make the data conform to the conceptual schema used by the warehouse
- Integrate the changes into the warehouse
  - Merge the data with existing data already present
  - Resolve possible update anomalies

Metadata Repository



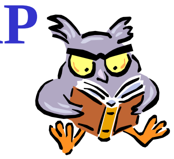
- Administrative metadata
  - Source database and their contents
  - Gateway descriptions
  - Warehouse schema, view and derived data definitions
  - Dimensions and hierarchies
  - Pre-defined queries and reports
  - Data mart locations and contents
  - Data partitions
  - Data extraction, cleansing, transformation rules, defaults
  - Data refresh and purge rules
  - User profiles, user groups
  - Security: user authorization, access control

Metadata Repository



- Business data
  - business terms and definitions
  - ownership of data
  - charging policies
- Operational metadata
  - data lineage: history of migrated data and sequence of transformations applied
  - currency of data: active, archived, purged
  - monitoring information: warehouse usage statistics, error reports, audit trails

Data Warehouse and OLAP Outline



- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?

Data Warehouse Design

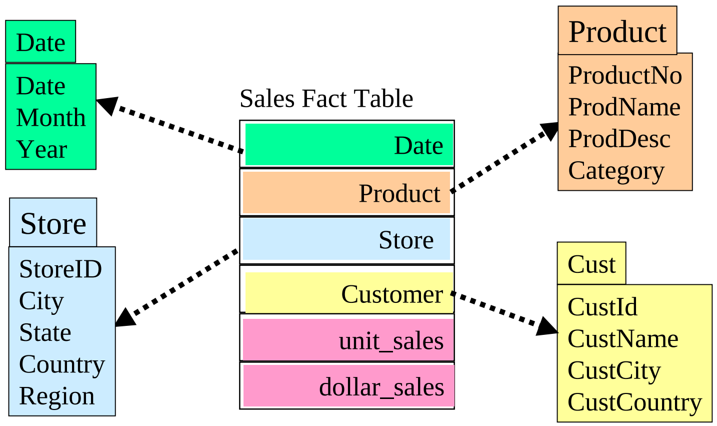
Most data warehouses use a **star schema** to represent the multi-dimensional model.

Each dimension is represented by a **dimension-table** that describes it.

A **fact-table** connects to all dimension-tables with a multiple join. Each tuple in the fact-table consists of a pointer to each of the dimension-tables that provide its multi-dimensional coordinates and stores measures for those coordinates.

The links between the fact-table in the centre and the dimension-tables in the extremities form a shape like a star. (*Star Schema*)

Example of Star Schema

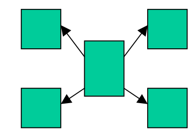


(Source: JH)

Data Warehouses Design (con't)

- Modeling data warehouses: dimensions & measurements

  Star schema: A single object (fact table) in the middle connected to a number of objects (dimension tables)



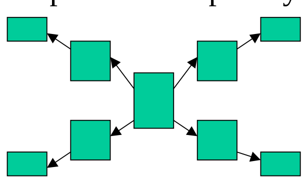
  Each dimension is represented by one table
  ➔Un-normalized (introduces redundancy).

  Ex: (Edmonton, Alberta, Canada, North America)
  (Calgary, Alberta, Canada, North America)

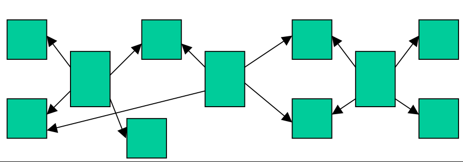
  Normalize dimension tables ➔ **Snowflake schema**

Data Warehouses Design (con't)

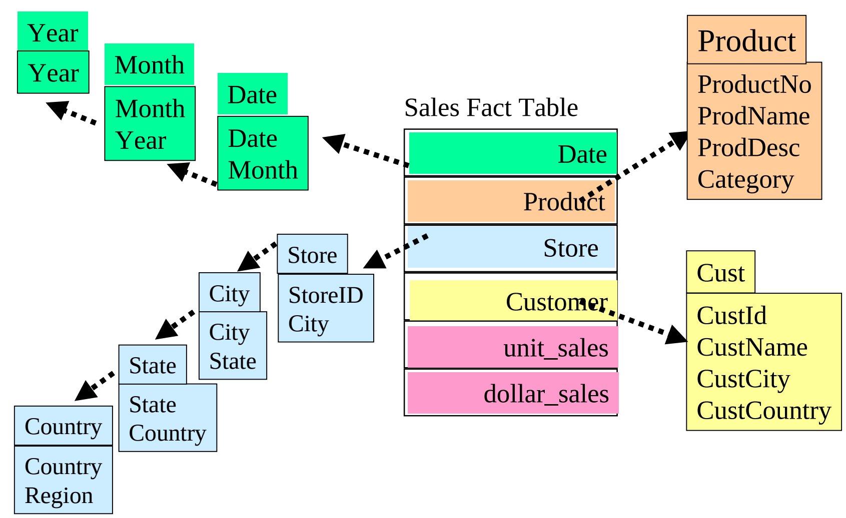
- **Snowflake schema**: A refinement of star schema where the dimensional hierarchy is represented explicitly by normalizing the dimension tables.



- **Fact constellations**: Multiple fact tables share dimension tables.



Example of Snowflake Schema



(Source: JH)

What Is the Best Design?

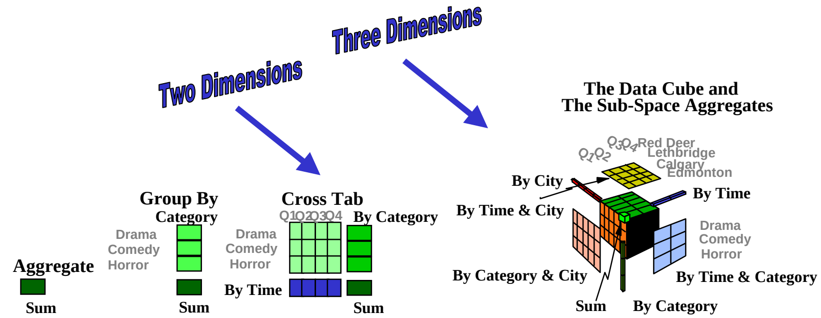
Performance benchmarking can be used to determine what is the best design.

Snowflake schema: Easier to maintain dimension tables when dimension table are very large (reduces overall space).

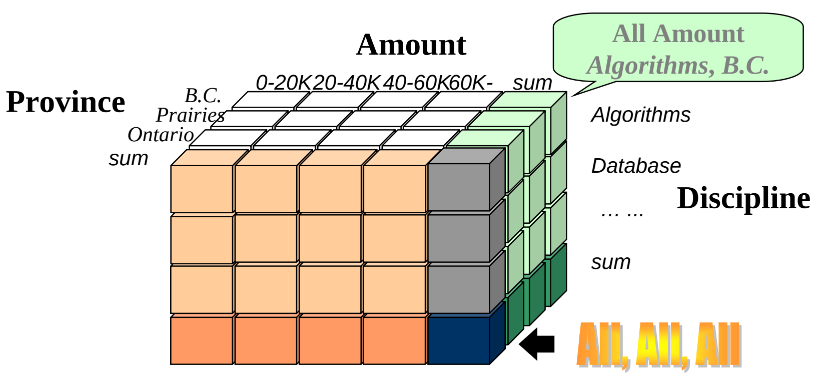
Star schema: More effective for data cube browsing (less joins): can affect performance.

Aggregation in Data Warehouses

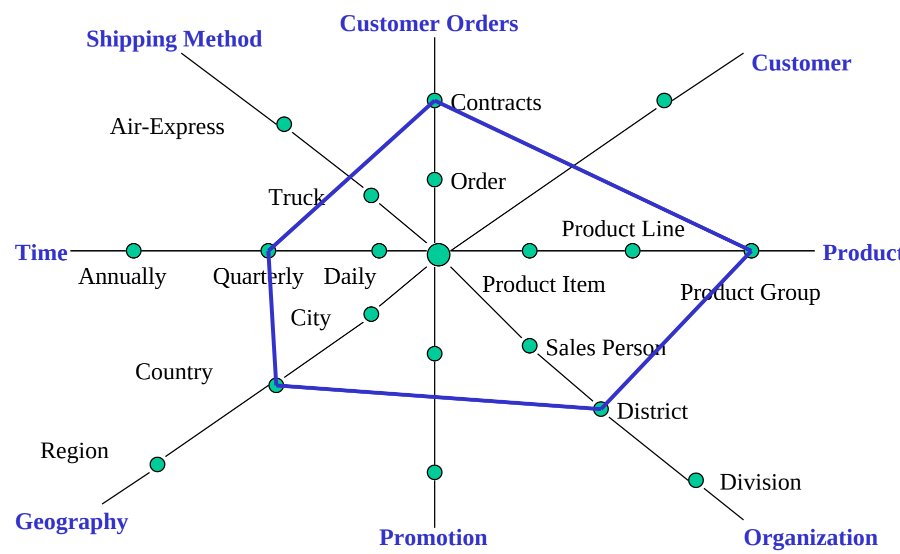
Multidimensional view of data in the warehouse:
Stress on aggregation of measures by one or more dimensions



Construction of Multi-dimensional Data Cube

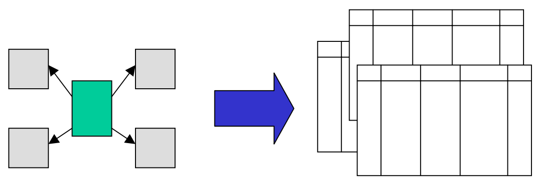


A Star-Net Query Model



Implementation of the OLAP Server

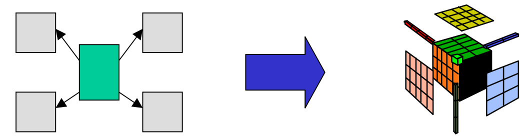
ROLAP: Relational OLAP - data is stored in tables in relational database or extended-relational databases. They use an RDBMS to manage the warehouse data and aggregations using often a star schema.
•They support extensions to SQL
•A cell in the multi-dimensional structure is represented by a tuple.
Advantage: Scalable (no empty cells for sparse cube).
Disadvantage: no direct access to cells.



Ex: Microstrategy Metacube (Informix)

Implementation of the OLAP Server

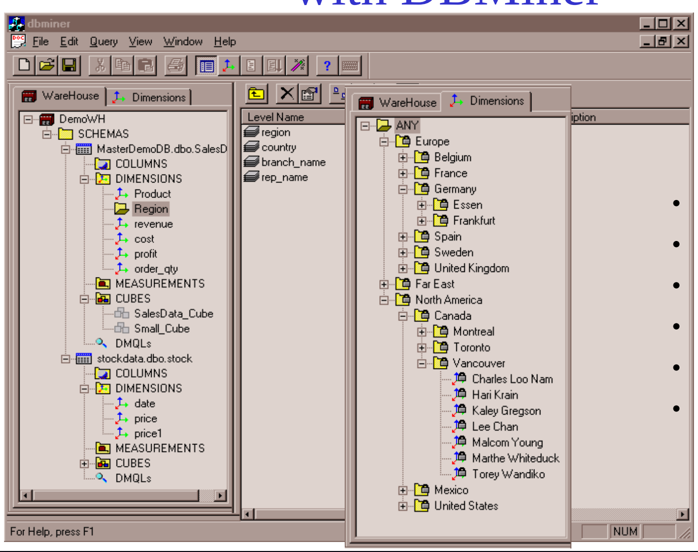
MOLAP: Multidimensional OLAP – implements the multidimensional view by storing data in special multidimensional data structures (MDDS)
Advantage: Fast indexing to pre-computed aggregations. Only values are stored.
Disadvantage: Not very scalable and sparse



Ex: Essbase of Arbor

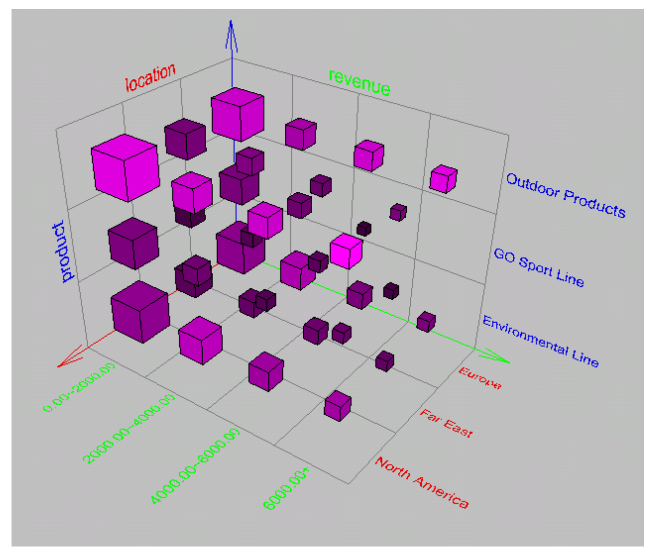
HOLAP: Hybrid OLAP - combines ROLAP and MOLAP technology. (Scalability of ROLAP and faster computation of MOLAP)

View of Warehouses and Hierarchies with DBMiner



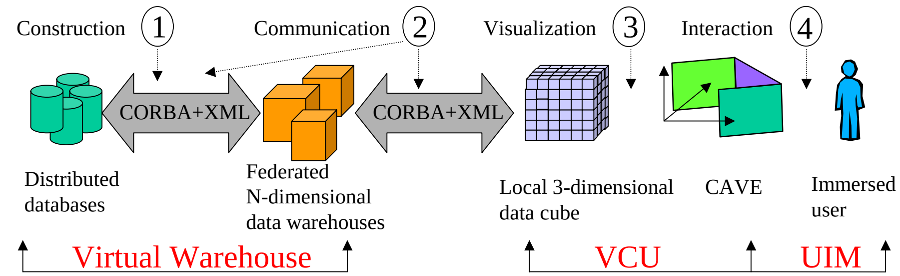
- Importing data
- Table Browsing
- Dimension creation
- Dimension browsing
- Cube building
- Cube browsing

DBMiner Cube Visualization

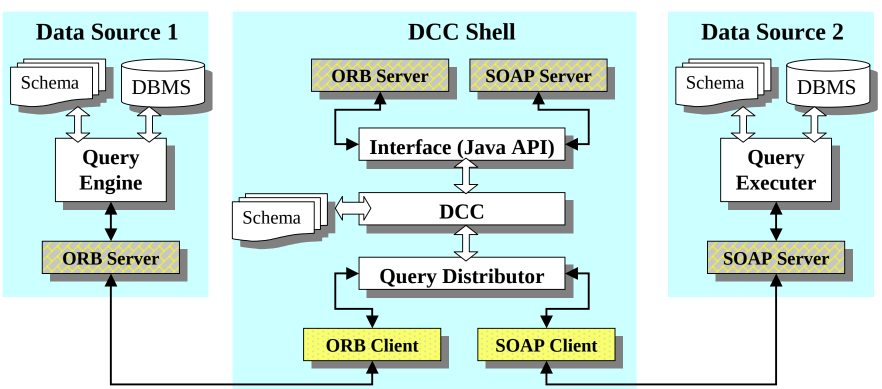


Example DIVE-ON Project

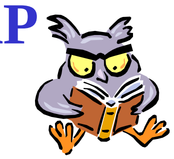
Example DIVE-ON Project



Example DIVE-ON Project

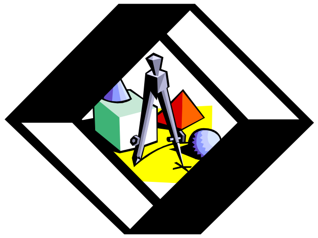


Data Warehouse and OLAP Outline



- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?

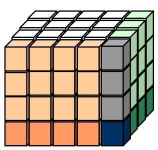
Issues



- Scalability
- Sparseness
- Curse of dimensionality
- Materialization of the multidimensional data cube (total, virtual, partial)
- Efficient computation of aggregations
- Indexing

Data Warehouse and OLAP Outline

- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?



Data Mining

- Data mining requires integrated, consistent and cleaned data which data warehouses can provide.
- Data mining tools can interface with the OLAP engine to take advantage of the integrated and aggregated data, as well as the navigation power.
- Interactive and exploratory mining.
- OLAP-based mining is referred to as OLAP-mining or OLAM (on-line analytical mining).