

CSE5334 DATA MINING

Lecture 4: Data Warehousing, OLAP, Data Cube

CSE 4392/5334 Data Mining, Fall 2009
Department of Computer Science and Engineering, University of Texas at Arlington
Chengkai Li (Slides courtesy of Jiawei Han)

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation

Lecture 4: Data Warehousing, OLAP

CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 2

What is Data Warehouse?

- "A data warehouse is a **subject-oriented, integrated, time-variant, and nonvolatile** collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing: The process of constructing and using data warehouses

Lecture 4: Data Warehousing, OLAP

CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 3

Data Warehouse—Subject-Oriented

- Organized around major subjects, such as **customer, product, sales**
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a **simple and concise** view around particular subject issues by **excluding data that are not useful in the decision support process**

Lecture 4: Data Warehousing, OLAP

CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 4

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
 - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
 - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
 - E.g., Hotel price: currency, tax, breakfast covered, etc.
 - When data is moved to the warehouse, it is converted.

Lecture 4: Data Warehousing, OLAP

CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 5

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
 - Operational database: current value data
 - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
 - Contains an element of time, explicitly or implicitly
 - But the key of operational data may or may not contain "time element"

Lecture 4: Data Warehousing, OLAP

CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 6

Data Warehouse—Nonvolatile

- A **physically separate store** of data transformed from the operational environment
- Operational **update of data does not occur** in the data warehouse environment
 - Does not require transaction processing, recovery, and concurrency control mechanisms
 - Requires only two operations in data accessing:
 - **initial loading of data and access of data**

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 7

Data Warehouse vs. Operational DBMS

- **OLTP (on-line transaction processing)**
 - Major task of traditional relational DBMS
 - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- **OLAP (on-line analytical processing)**
 - Major task of data warehouse system
 - Data analysis and decision making
- **Distinct features (OLTP vs. OLAP):**
 - User and system orientation: customer vs. market
 - Data contents: current, detailed vs. historical, consolidated
 - Database designs: ER + application vs. star + subject
 - View: current, local vs. evolutionary, integrated
 - Access patterns: update vs. read-only but complex queries

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 8

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

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 9

Why Separate Data Warehouse?

- Different functions and different data:
- **Note:** There are more and more systems which perform OLAP analysis directly on relational databases
- There is no absolute boundary.

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 10

Chapter 3: Data Warehousing and OLAP Technology: An Overview

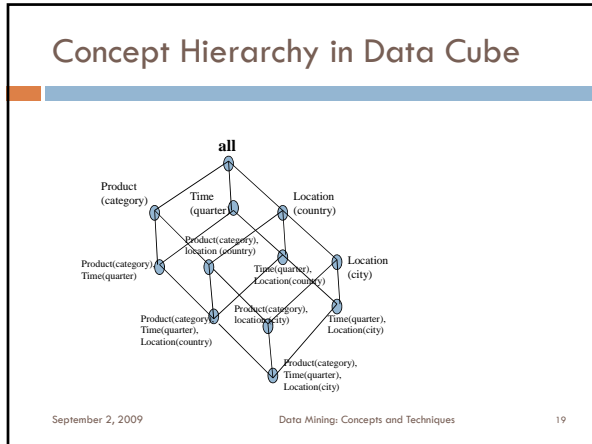
- What is a data warehouse?
- A **multi-dimensional data model**
- Data warehouse architecture
- Data warehouse implementation

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 11

Data Cube

- A data warehouse is based on a **multidimensional data model** which views data in the form of a data cube
- A data cube contains **aggregates of measure values**, on various combinations of dimensions, and furthermore, with various levels of aggregation on individual dimension.
- In data warehousing literature, an n-D base cube is called a **base cuboid**. The top most 0-D cuboid, which holds the highest-level of summarization, is called the **apex cuboid**. The lattice of cuboids forms a **data cube**.

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 12



Conceptual Schema Design

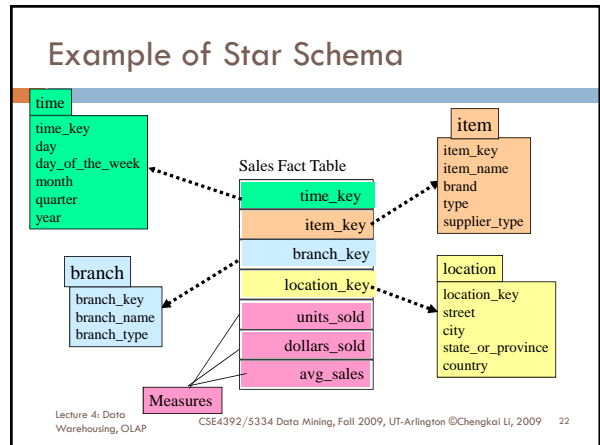
- **Dimensions & Measures**
 - **Dimension tables**, such as product (item_name, brand, type), or time(day, week, month, quarter, year)
 - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 20

Conceptual Modeling of Data Warehouses

- **Star schema**: A fact table in the middle connected to a set of dimension tables

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 21



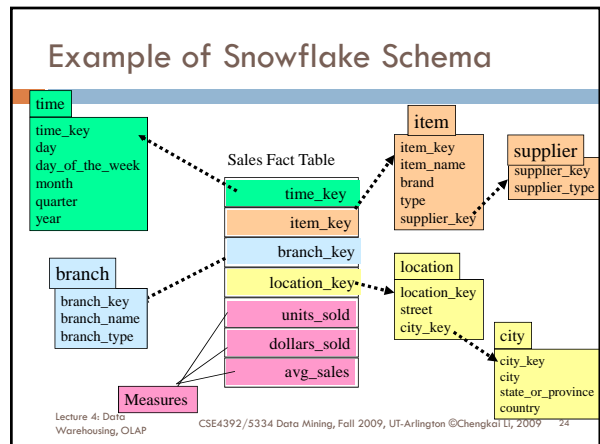
Conceptual Modeling of Data Warehouses

- **Snowflake schema**: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake

It provides explicit support of hierarchy

- Easier to manage the dimension
- Can be less efficient (due to join) than star schema

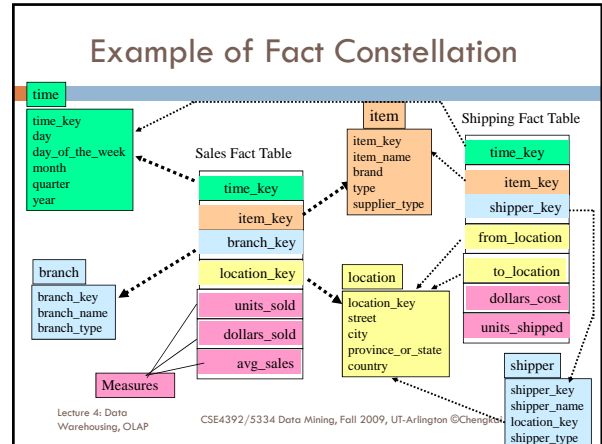
Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 23



Conceptual Modeling of Data Warehouses

- Fact constellations:** Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called **galaxy schema** or fact constellation

Lecture 4: Data Warehousing, OLAP
CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 25



Measures of Data Cube: Three Categories

- Distributive:** if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning
 - E.g., count(), sum(), min(), max()
- Algebraic:** if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function
 - E.g., avg(), min_N(), standard_deviation()
- Holistic:** if there is no constant bound on the storage size needed to describe a subaggregate.
 - E.g., median(), mode(), rank()

Lecture 4: Data Warehousing, OLAP
CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 27

Typical OLAP Operations

- Roll up (drill-up):** summarize data
 - by climbing up hierarchy or by dimension reduction
- Drill down (roll down):** reverse of roll-up
 - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- Slice and dice:** project and select
- Pivot (rotate):**
 - reorient the cube, visualization, 3D to series of 2D planes

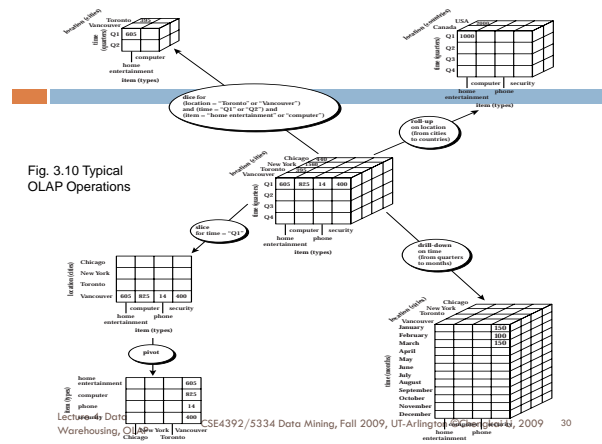
Lecture 4: Data Warehousing, OLAP
CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 28

Roll up and Drill Down

- Roll up: increasing the level of aggregation**
 - further aggregating along one more dimension
 - or further aggregating along the hierarchy of one dimension
- Drill down: decreasing the level of aggregating**

It is like traversing in the lattice of cuboids.

Lecture 4: Data Warehousing, OLAP
CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 29



Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 31

OLAP Server Architectures

- **Relational OLAP (ROLAP)**
 - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware
 - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
 - Greater scalability
- **Multidimensional OLAP (MOLAP)**
 - Sparse array-based multidimensional storage engine
 - Fast indexing to pre-computed summarized data
- **Hybrid OLAP (HOLAP)** (e.g., Microsoft SQLServer)
 - Flexibility, e.g., low level: relational, high-level: array
- **Specialized SQL servers** (e.g., Redbricks)
 - Specialized support for SQL queries over star/snowflake schemas

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 32

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 33

Efficient Data Cube Computation

- Data cube can be viewed as a lattice of cuboids
 - The bottom-most cuboid is the base cuboid
 - The top-most cuboid (apex) contains only one cell
 - How many cuboids in an n-dimensional cube with L levels?

$$T = \prod_{i=1}^n (L_i + 1)$$

- Materialization of data cube
 - Materialize **every** (cuboid) (full materialization), **none** (no materialization), or **some** (partial materialization)
 - Selection of which cuboids to materialize
 - Based on size, sharing, access frequency, etc.

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 34

Cube Operation

- Cube definition and computation in DMQL


```
define cube sales[item, city, year]: sum(sales_in_dollars)
compute cube sales
```
- Transform it into a SQL-like language (with a new operator **cube by**, introduced by Gray et al.'96)


```
SELECT item, city, year, SUM (amount)
FROM SALES
CUBE BY item, city, year
```
- Need to compute the following Group-Bys


```
(date, product, customer),
(date,product),(date, customer), (product, customer),
(date), (product), (customer)
()
```

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 35

Iceberg Cube

- Computing only the cuboid cells whose count or other aggregates satisfying the condition like


```
HAVING COUNT(*) >= minsup
```
- Motivation
 - Only a small portion of cube cells may be “above the water” in a sparse cube
 - Only calculate “interesting” cells—data above certain threshold
 - Avoid explosive growth of the cube
 - Suppose 100 dimensions, only 1 base cell. How many aggregate cells if count >= 1? What about count >= 2?

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 36

Indexing OLAP Data: Bitmap Index

- Index on a particular column
- Each value in the column has a bit vector: bit-op is fast
- The length of the bit vector: # of records in the base table
- The *i*-th bit is set if the *i*-th row of the base table has the value for the indexed column
- not suitable for high cardinality domains

Base table			Index on Region			Index on Type			
Cust	Region	Type	RecID	Asia	Europe	America	RecID	Retail	Dealer
C1	Asia	Retail	1	1	0	0	1	1	0
C2	Europe	Dealer	2	0	1	0	2	0	1
C3	Asia	Dealer	3	1	0	0	3	0	1
C4	America	Retail	4	0	0	1	4	1	0
C5	Europe	Dealer	5	0	1	0	5	0	1

Indexing OLAP Data: Join Indices

- Join index: $J(R-id, S-id)$ where $R(R-id, \dots) \triangleright \triangleleft S(S-id, \dots)$
- Traditional indices map the values to a list of record ids
- It materializes relational join in *JI* file and speeds up relational join
- In data warehouses, join index relates the values of the **dimensions** of a start schema to **rows** in the fact table.
 - E.g. fact table: *Sales* and two dimensions *city* and *product*
 - A join index on *city* maintains for each distinct *city* a list of R-IDs of the tuples recording the *Sales* in the *city*
 - Join indices can span multiple dimensions

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 38

Efficient Processing OLAP Queries

- Determine which operations should be performed on the available cuboids
 - Transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g., dice = selection + projection
- Determine which materialized cuboid(s) should be selected for OLAP op.
 - Let the query to be processed be on {brand, province_or_state} with the condition "year = 2004", and there are 4 materialized cuboids available:
 - 1) {year, item_name, city}
 - 2) {year, brand, country}
 - 3) {year, brand, province_or_state}
 - 4) {item_name, province_or_state} where year = 2004
 Which should be selected to process the query?
- Explore indexing structures and compressed vs. dense array structures in MOLAP

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation
- Summary

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 40

Summary: Data Warehouse and OLAP Technology

- Why data warehousing?
- A **multi-dimensional model** of a data warehouse
 - Star schema, snowflake schema, fact constellations
 - A data cube consists of dimensions & measures
- **OLAP** operations: drilling, rolling, slicing, dicing and pivoting
- Data warehouse architecture
 - OLAP servers: ROLAP, MOLAP, HOLAP
- Efficient computation of data cubes
 - Partial vs. full vs. no materialization
 - Indexing OALP data: Bitmap index and join index
 - OLAP query processing

Lecture 4: Data Warehousing, OLAP CSE4392/5334 Data Mining, Fall 2009, UT-Arlington ©Chengkai Li, 2009 41