Chapter IV: OLAP

Knowledge Discovery in Databases

Luciano Melodia M.A.
Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg
Summer semester 2021

Chapter IV: Data warehousing and online analytical processing

Data warehouse: basic concepts.

Data-warehouse modeling: data cube and OLAP.

Data-warehouse design and usage.

Data-warehouse Implementation.

Data generalization by attribute-oriented induction.

Summary.

What is a data warehouse?

Defined in many different ways, but not rigorously:

A **decision-support** database that is **maintained separately** from the organization's operational database.
Supports information processing by providing a solid platform of **consolidated, historical data** for analysis.

Famous:
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.

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.

Data warehouse – integrated

Constructed by integrating multiple heterogeneous data sources.

Relational databases, flat files, online 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.
ETL – Extraction, Transformation, Loading, see below.

Data warehouse – time variant

The time horizon for a data warehouse is significantly longer than that of operational systems.

Operational database: current-value data.

Data warehouse: 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.

The key of operational data may or may not contain a "time element."

Data warehouse – nonvolatile

A physically separate store of data.

Transformed from the operational environment.
By copying.

No operational update of data:

Hence, does not require transaction processing,
i.e. no logging, recovery, concurrency control, etc.
Requires only three operations:

Initial loading of data.
Refresh (update, often periodically, e.g. over night).
Access of data.

OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day-to-day operations	decision support
DB design	application-oriented	decision support
data	current, up-to-date; detailed, flat relational; isolated	historical; summarized, multidimensional, integrated, consolidated
usage	repetitive	ad-hoc
access	read/write; index/hash on primary key	lots of scans
unit of work	short, simple transaction	complex query
#-records accessed	10	10^6
#-users	1000	100
DB size	100 MB to GB	100 GB to TB
quantification	transaction throughput	query throughput, response

Why a separate data warehouse?

High performance for both systems:

- **DBMS**: tuned for OLTP; Access methods, indexing concurreny control, recovery.
- **Warehouse**: tuned for OLAP; Complex OLAP queries, multidimensional view, consolidation.

Different functions and different data:

- Missing data:
  - Decision support (DS) requires **historical data** which operational DBs do not typically maintain.
- Data consolidation:
  - DS requires **consolidation** (aggregation, summarization) of data from heterogeneous sources.
- Data quality:
  - Different sources typically use inconsistent data representations, codes and formats which have to be reconciled.

Note: There are more and more systems which perform OLAP analysis directly on relational databases.

References (I)

S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi: On the computation of multidimensional aggregates. VLDB'96.

D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek: Efficient view maintenance in data warehouses. SIGMOD'97.

R. Agrawal, A. Gupta, and S. Sarawagi: Modeling multidimensional databases. ICDE'97.

S. Chaudhuri and U. Dayal: An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997-

E. F. Codd, S. B. Codd, and C. T. Salley: Beyond decision support. Computer World, 27, July 1993.

J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.

A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999.

References (II)

J. Han: Towards on-line analytical mining in large databases. ACM SIGMOD Record, 27:97-107, 1998.

V. Harinarayan, A. Rajaraman, and J. D. Ullman: Implementing data cubes efficiently. SIGMOD'96.

J. Hellerstein, P. Haas, and H. Wang: Online aggregation. SIGMOD'97.

C. Imhoff, N. Galemmo, and J. G. Geiger: Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003.

W. H. Inmon: Building the Data Warehouse. John Wiley, 1996.

R. Kimball and M. Ross: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002.

P. O'Neil and G. Graefe: Multi-table joins through bitmapped join indices. ACM SIGMOD Record, 24:8–11, Sept. 1995.

P. O'Neil and D. Quass: Improved query performance with variant indexes. SIGMOD'97.

Microsoft. OLEDB for OLAP programmer's reference version 1.0. In http://www.microsoft.com/data/oledb/olap, 1998.

Thank you for your attention.
Any questions about the fourth chapter?

Ask them now, or again, drop me a line:
luciano.melodia@fau.de.