(c) Suppose the function is "*top 10 sales*". Discuss how to efficiently compute this measure in a data cube.

7. Suppose that one needs to record three measures in a data cube: `min`, `average`, and `median`. Design an efficient computation and storage method for each measure given that the cube allows data to be *deleted* incrementally (i.e., in small portions at a time) from the cube.

8. In data warehouse technology, a multiple dimensional view can be implemented by a multidimensional database technique (MOLAP), or by a relational database technique (ROLAP), or a hybrid database technique (HOLAP).

   (a) Briefly describe each implementation technique.

   (b) For each technique, explain how each of the following functions may be implemented:

      i. The generation of a data warehouse (including aggregation).
      ii. Roll-up.
      iii. Drill-down.
      iv. Incremental updating.

   Which implementation techniques do you prefer, and why?

9. Suppose that a data warehouse contains 20 dimensions each with about 5 levels of granularity.

   (a) Users are mainly interested in four particular dimensions, each having three frequently accessed levels for rolling up and drilling down. How would you design a data cube structure to support this preference efficiently?

   (b) At times, a user may want to *drill-through* the cube, down to the raw data for one or two particular dimensions. How would you support this feature?

10. Data cube computation: Suppose a base cuboid has 3 dimensions, $(A, B, C)$, with the number of cells shown below: $|A| = 1,000,000$, $|B| = 100$, and $|C| = 1,000$. Suppose each dimension is partitioned evenly into 10 portions for chunking.

    (a) Assuming each dimension has only one level, draw the complete lattice of the cube.

    (b) If each cube cell stores one measure with 4 bytes, what is the total size of the computed cube if the cube is *dense*?

    (c) If the cube is very *sparse*, describe an effective multidimensional array structure to store the sparse cube.

    (d) State the order for computing the chunks in the cube which requires the least amount of space, and compute the total amount of main memory space required for computing the 2-D planes.

11. In both data warehousing and data mining, it is important to have some hierarchy information associated with each dimension. If such a hierarchy is not given, discuss how to generate such a hierarchy automatically for the following cases:

    (a) a dimension containing only numerical data.

    (b) a dimension containing only categorical data.

12. Suppose that a data cube has 2 dimensions, $(A, B)$, and each dimension can be generalized through 3 levels (with the top-most level being all). That is, starting with level $A_0$, $A$ can be generalized to $A_1$, then to $A_2$, and then to all. How many *different* cuboids (i.e., views) can be generated for this cube? Sketch a lattice of these cuboids to show how you derive your answer. Also, give a general formula for a data cube with $D$ dimensions, each starting at a base level and going up through $L$ levels, with the top-most level being all.

13. Consider the following *multifeature cube* query: Grouping by all subsets of {item, region, month}, find the minimum shelf life in 1997 for each group, and the fraction of the total sales due to tuples whose price is less than \$100, and whose shelf life is within 25% of the minimum shelf life, and within 50% of the minimum shelf life.

    (a) Draw the multifeature cube graph for the query.

(b) Express the query in extended SQL.

(c) Is this a *distributive* multifeature cube? Why or why not?

14. What are the differences between the three main types of data warehouse usage: *information processing*, *analytical processing*, and *data mining*? Discuss the motivation behind *OLAP mining (OLAM)*.

## Bibliographic Notes

There are a good number of introductory level textbooks on data warehousing and OLAP technology, including Inmon [15], Kimball [16], Berson and Smith [4], and Thomsen [24]. Chaudhuri and Dayal [6] provide a general overview of data warehousing and OLAP technology.

The history of decision support systems can be traced back to the 1960s. However, the proposal of the construction of large data warehouses for multidimensional data analysis is credited to Codd [7] who coined the term *OLAP* for *on-line analytical processing*. The OLAP council was established in 1995. Widom [26] identified several research problems in data warehousing. Kimball [16] provides an overview of the deficiencies of SQL regarding the ability to support comparisons that are common in the business world.

The DMQL data mining query language was proposed by Han et al. [11] Data mining query languages are further discussed in Chapter 4. Other SQL-based languages for data mining are proposed in Imielinski, Virmani, and Abdulghani [14], Meo, Psaila, and Ceri [17], and Baralis and Psaila [3].

Gray et al. [9, 10] proposed the data cube as a relational aggregation operator generalizing group-by, crosstabs, and sub-totals. Harinarayan, Rajaraman, and Ullman [13] proposed a greedy algorithm for the partial materialization of cuboids in the computation of a data cube. Agarwal et al. [1] proposed several methods for the efficient computation of multidimensional aggregates for ROLAP servers. The chunk-based multiway array aggregation method described in Section 2.4.1 for data cube computation in MOLAP was proposed in Zhao, Deshpande, and Naughton [27]. Additional methods for the fast computation of data cubes can be found in Beyer and Ramakrishnan [5], and Ross and Srivastava [19]. Sarawagi and Stonebraker [22] developed a chunk-based computation technique for the efficient organization of large multidimensional arrays.

For work on the selection of materialized cuboids for efficient OLAP query processing, see Harinarayan, Rajaraman, and Ullman [13], and Sristava et al. [23]. Methods for cube size estimation can be found in Beyer and Ramakrishnan [5], Ross and Srivastava [19], and Deshpande et al. [8]. Agrawal, Gupta, and Sarawagi [2] proposed operations for modeling multidimensional databases.

The use of join indices to speed up relational query processing was proposed by Valduriez [25]. O'Neil and Graefe [18] proposed a bitmapped join index method to speed-up OLAP-based query processing.

There are some recent studies on the implementation of discovery-oriented data cubes for data mining. This includes the discovery-driven exploration of OLAP data cubes by Sarawagi, Agrawal, and Megiddo [21], and the construction of multifeature data cubes by Ross, Srivastava, and Chatziantoniou [20]. For a discussion of methodologies for OLAM (On-Line Analytical Mining), see Han et al. [12].

# Bibliography


[1] S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proc. 1996 Int. Conf. Very Large Data Bases*, pages 506–521, Bombay, India, Sept. 1996.

[2] R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. In *Proc. 1997 Int. Conf. Data Engineering*, pages 232–243, Birmingham, England, April 1997.

[3] E. Baralis and G. Psaila. Designing templates for mining association rules. *Journal of Intelligent Information Systems*, 9:7–32, 1997.

[4] A. Berson and S. J. Smith. *Data Warehousing, Data Mining, and OLAP*. New York: McGraw-Hill, 1997.

[5] K. Beyer and R. Ramakrishnan. Bottom-up computation of sparse and iceberg cubes. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data*, pages 359–370, Philadelphia, PA, June 1999.

[6] S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26:65–74, 1997.

[7] E. F Codd, S. B. Codd, and C. T. Salley. Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. In *E. F. Codd & Associates available at http://www.arborsoft.com/OLAP.html*, 1993.

[8] P. Deshpande, J. Naughton, K. Ramasamy, A. Shukla, K. Tufte, and Y. Zhao. Cubing algorithms, storage estimation, and storage and processing alternatives for olap. *Data Engineering Bulletin*, 20:3–11, 1997.

[9] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: A relational operator generalizing group-by, cross-tab and sub-totals. In *Proc. 1996 Int. Conf. Data Engineering*, pages 152–159, New Orleans, Louisiana, Feb. 1996.

[10] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Mining and Knowledge Discovery*, 1:29–54, 1997.

[11] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. R. Zaïane. DBMiner: A system for mining knowledge in large relational databases. In *Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 250–255, Portland, Oregon, August 1996.

[12] J. Han, Y. J. Tam, E. Kim, H. Zhu, and S. H. S. Chee. Methodologies for integration of data mining and on-line analytical processing in data warehouses. In *submitted to DAMI*, 1999.

[13] V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, pages 205–216, Montreal, Canada, June 1996.

[14] T. Imielinski, A. Virmani, and A. Abdulghani. DataMine – application programming interface and query language for KDD applications. In *Proc. 1996 Int. Conf. Data Mining and Knowledge Discovery (KDD'96)*, pages 256–261, Portland, Oregon, August 1996.

[15] W. H. Inmon. *Building the Data Warehouse*. QED Technical Publishing Group, Wellesley, Massachusetts, 1992.

[16] R. Kimball. *The Data Warehouse Toolkit*. John Wiley & Sons, New York, 1996.

[17] R. Meo, G. Psaila, and S. Ceri. A new SQL-like operator for mining association rules. In *Proc. 1996 Int. Conf. Very Large Data Bases*, pages 122–133, Bombay, India, Sept. 1996.

[18] P. O'Neil and G. Graefe. Multi-table joins through bitmapped join indices. *SIGMOD Record*, 24:8–11, September 1995.

[19] K. Ross and D. Srivastava. Fast computation of sparse datacubes. In *Proc. 1997 Int. Conf. Very Large Data Bases*, pages 116–125, Athens, Greece, Aug. 1997.

[20] K. A. Ross, D. Srivastava, and D. Chatziantoniou. Complex aggregation at multiple granularities. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98)*, pages 263–277, Valencia, Spain, March 1998.

[21] S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. In *Proc. Int. Conf. of Extending Database Technology (EDBT'98)*, pages 168–182, Valencia, Spain, March 1998.

[22] S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. In *Proc. 1994 Int. Conf. Data Engineering*, pages 328–336, Feb. 1994.

[23] D. Sristava, S. Dar, H. V. Jagadish, and A. V. Levy. Answering queries with aggregation using views. In *Proc. 1996 Int. Conf. Very Large Data Bases*, pages 318–329, Bombay, India, September 1996.

[24] E. Thomsen. *OLAP Solutions: Building Multidimensional Information Systems.* John Wiley & Sons, 1997.

[25] P. Valduriez. Join indices. In *ACM Trans. Database System*, volume 12, pages 218–246, 1987.

[26] J. Widom. Research problems in data warehousing. In *Proc. 4th Int. Conf. Information and Knowledge Management*, pages 25–30, Baltimore, Maryland, Nov. 1995.

[27] Y. Zhao, P. M. Deshpande, and J. F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. In *Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data*, pages 159–170, Tucson, Arizona, May 1997.