# ADVANCED TECHNIQUES FOR SCIENTIFIC DATAWAREHOUSES

**S. Sai Satyanarayana Reddy[1] , Dr.L.S.S.Reddy[2], Dr.V.Khanaa[3], A.Lavanya[4],**

[1]*Reseach Scholar* [2]*internal Guide &Principal (KLC)* [3]*Supervosor& Guide* [4] *Student*

[1] *PadmaSri Dr. B.V.Raju Institute of Technology, Vishnu Pur, Narsapur, Medhak (Dt), A.P, India*

***Email**: saisn90@gmail.com, meetss90@yahoo.in, phani_lav@yahoo.co.in*

***Phone**: +91-9440012540*

**Abstract**

*Data warehouses using a multidimensional view of data have become very popular in both business and science in recent years. Data warehouses for scientific purposes such as medicine and bio-chemistry1 pose several great challenges to existing data warehouse technology. Data warehouses usually use pre-aggregated data to ensure fast query response. However, pre-aggregation cannot be used in practice if the dimension structures or the relationships between facts and dimensions are irregular. A technique for overcoming this limitation and some experimental results are presented. Queries over scientific data warehouses often need to reference data that is external to the data warehouse, e.g., data that is too complex to be handled by current data warehouse technology, data that is "owned" by other organizations, or data that is updated frequently. This paper presents a federation architecture that allows the integration of multidimensional warehouse data with complex external data.*

## 1. INTRODUCTION

Data Warehousing (DW) and On-Line Analytical Processing (OLAP) systems based on a dimensional view of data are being used increasingly in traditional business applications as well as in applications such as health care and bio-chemistry for the purpose of analyzing very large amounts of data. The use of DW and OLAP systems for scientific purposes raises several new challenges to the traditional technology. This paper describes two of these challenges, both of which are concerned with implementation aspects. The first is the optimal use of pre-aggregated data for improved query performance even when the data structures are irregular, while the second is the integration of multidimensional OLAP databases with complex external data. Other challenges are related to the conceptual and logical design of scientific data warehouses, including modeling and querying complex multidimensional data and handling imprecise data. However, these challenges are beyond the scope of this paper. In order to improve query performance, modern OLAP systems use a technique known as practical pre-aggregation, where combinations of aggregate queries are materialized selectively and re-used when computing other aggregates; full pre-aggregation, where all combinations of aggregates are materialized, is infeasible, as it typically causes a blowup in storage requirements of 200–500 times the size of the raw data. Normally, practical pre-aggregation requires the dimension hierarchies to be regular, i.e., to be balanced trees, but this is quite often not the case in real-world systems. The technique presented here enables practical pre-aggregation even for irregular hierarchies. The details of the technique can be found elsewhere. We show how to achieve practical pre-aggregation through transformations of the dimensions and how the transformations can be accomplished transparently to the user. The technique enables the achievement of fast query response time while saving huge amounts of storage compared to current OLAP systems and techniques. The prototype implementation of TreeScape demonstrates that these benefits may be achieved with standard technology.

### 2.2 OLAP + Data Mining → On-Line Analytical Mining

On-line analytical processing (OLAP) is a powerful data analysis method for multi-dimensional analysis of data warehouses. Motivated by the popularity of OLAP technology, we develop an On-Line Analytical Mining (OLAM) mechanism for multi-dimensional data mining in large databases and data warehouses. We believe this is a promising direction to pursue based on the following observations.

1. Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data transformation and data integration as preprocessing steps. A data warehouse constructed by such preprocessing serves as a valuable source of cleaned and integrated data for OLAP as well as for data mining.
2. Effective data mining needs exploratory data analysis. A users often likes to traverse flexibly through a database, select any portions of relevant data, analyze data at different granularities, and present knowledge/results in different forms. On-

978-0-7695-3516-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICACC.2009.145

line analytical mining provides facilities for data mining on different subsets of data and at different levels of abstraction, by drilling, pivoting, filtering, dicing and slicing on a data cube and on some intermediate data mining results. This, together with data/knowledge visualization tools, will greatly enhance the power and flexibility of exploratory data mining.

3. It is often difficult for a user to predict what kinds of knowledge to be mined beforehand. By integration of OLAP with multiple data mining functions. On-line analytical mining provides flexibility for users to select desired data mining functions and swap data mining tasks dynamically.

### 2.3 Architecture for on-line analytical mining

An OLAM engine performs analytical mining in data cubes in a similar manner as an OLAP engine performs on-line analytical processing. Therefore, it is suggested to have an integrated OLAM and OLAP architecture as shown in below Figure where the OLAM and OLAP engines both accept users' on-line queries (instructions) and work with the data cube in

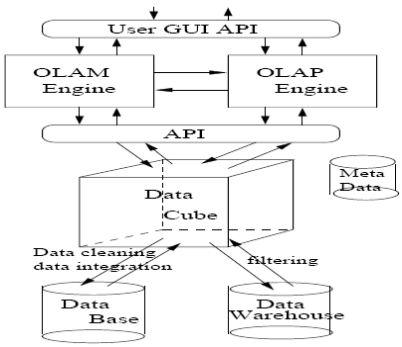



**Figure:** An integrated OLAM and OLAP architecture the analysis Furthermore, an OLAM engine may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, etc. Therefore, an OLAM engine is more sophisticated than an OLAP engine since it usually consists of multiple mining modules which may interact with each other for effective mining.

#### 2.3.1 Data cube construction

Data cube technology is essential for efficient on-line analytical mining. There have been many studies on efficient computation and access of multidimensional databases.

Our early development of attribute-oriented induction method adopts two generalization techniques (1) attribute removal, which removes attributes which represent low-level data in a hierarchy, and (2) attribute generalization which generalizes attribute values to their corresponding high level ones. Such generalization leads to a new, compressed generalized relation with count and/or other aggregate values accumulated. This is similar to the relational OLAP (ROLAP) implementation of the roll-up operation.

For fast response in OLAP and data mining, our later implementation has adopted data cube technology as follows, when data cube contains a small number of dimensions, or when it is generalized to a high level, the cube is structured as compressed sparse array but is still stored in a relational database (to reduce the cost of construction and indexing of different data structures).

We believe such a dual data structure technique represents a balance between multidimensional OLAP (MOLAP) and relational OLAP (ROLAP) implementations. It ensures fast response time when handling medium-sized cubes/cuboids and high scalability when handling large databases with high dimensionality.

Notice that even adopting the ROLAP technique, it is still unrealistic to materialize all the possible cuboids for large databases with high dimensionality due to the huge number of cuboids It is wise to materialize more of the generalized, low dimensionality cuboids besides considering other factors, such as accessing patterns and the sharing among different cuboids.

A 3-D data cube/cuboid can be selected from a high-dimensional data cube and be browsed conveniently using the DBMiner 3-D cube browser as shown in Figure, where the size of a cell (displayed as a tiny cube) represents the entry count in the corresponding cell, and the brightness of the cell represents another measure of the cell. Pivoting, drilling, and slicing/dicing operations can be performed on the data cube browser with mouse clicking.

#### 2.3.2 Concept description

Concept/class description plays an important role in descriptive data mining. It consists of two major functions, data characterization and data discrimination .

Data characterization summarizes and characterizes a set of task-relevant data by data generalization. Data characterization and its associated OLAP operations, such as drill-down, roll-up (also called drill-up),

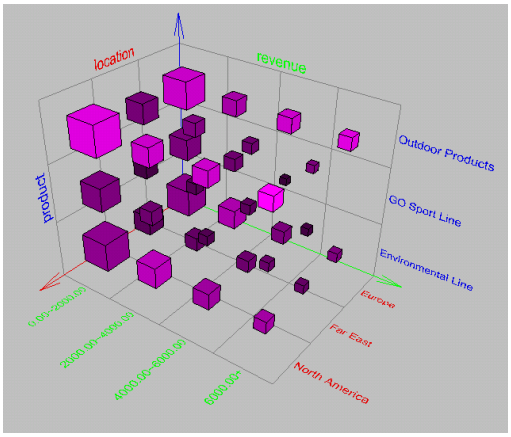



**Figure:** Browsing of a 3-dimensional data cube in DBMiner

slice, and dice can be performed on data cubes. Drilling operation facilitates users to examine data characteristics at multiple levels of abstraction.

## 3. OLAP++ SYSTEM ARCHITECTURE

The overall architecture of the OLAP++ system is seen in Figure. The object part of the system is based on the OPM tools that implements the Object Data Management Group (ODMG) object data model and the Object Query Language (OQL) on top of a relational DBMS, in this case the ORACLE RDBMS. The OLAP part of the system is based on Microsoft's SQL Server OLAP Services using the Multi-Dimensional eXpressions (MDX) query language.

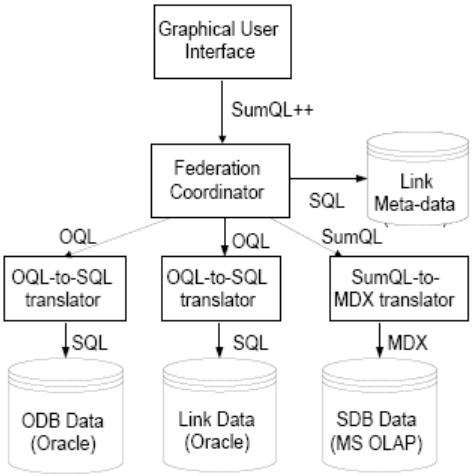



**Figure:** OLAP++ Architecture

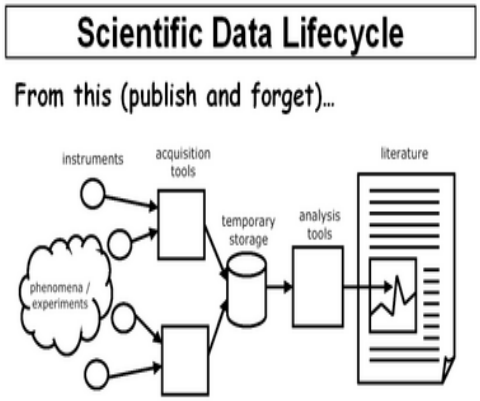
When a SumQL++ query is received by the Federation Coordinator (FC), it is first parsed to identify the measures, categories, links, classes and attributes referenced in the query. Based on this, the FC then queries the metadata to get information about which databases the object data and the OLAP data reside in and which categories are linked to which classes. Based on the object parts of the query, the FC then sends OQL queries to the object databases to retrieve the data for which the particular conditions holds true. This data is then put into a"pure" SumQL statement (i.e. without object references) as a list of category values. The SumQL statement is translated into MDX by a separate layer, the"SumQL-to-MDX translator", and the data returned from OLAP Services is returned to the FC. The reason for using the intermediate SumQL statements is to isolate the implementation of the OLAP data from the FC. As an another alternative, we have also implemented a translator into SQL statements against a"star schema" relational database design. The system is able to support a good query performance even for large databases while making it possible to integrate existing OLAP data with external data in object databases in a flexible way that can adapt quickly to changing query needs.

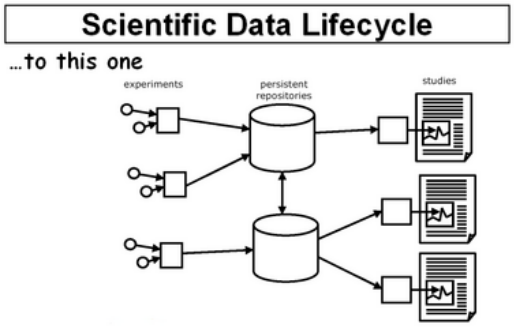
### Problems and opportunities

Problems that can be found in current scientific projects are for example:

- Limited file and directory naming schemes. Some project data repositories are simply big flat directories.
- Scientists retrieve entire files to ascertain relevance.
- No access to important metadata in scientists' notebooks and heads.
- Un-owned data with dubious content after the end of project or PhD thesis.

But the increasing of scientific data collections size brings not only problems, but also a lot of opportunities. One of the biggest opportunities is the possibility of reuse existing data for new studies. One example is provided by the various Virtual Observatory initiatives in Europe and USA. The idea is summarized below:






Another virtuous effect can be called "discovery by browsing". If the data is well described and the data access method is quite flexible, the user can establish unexpected correlations between data items thus facilitating serendipitous discoveries.

Last, but not least, remember that the data is composed not only by bytes, but also by workflow definitions, computation parameters, environment setup and so on.

**An important note**: the biopharmaceutical industry attacks a more specific meaning to Scientific Data Management. They also have huge data sets to be managed, but they must comply also with industry regulations and rigidly enforce intellectual property protection. The second point is important for each science field, but not as vital as in industry. In this paper we don't touch those specific problems.

## 4. APPPLICATIONS

### 4.1 Medical microrobotics

There are ongoing attempts to build microrobots for in vivo medical use. In 2002, Ishiyama et al. at Tohoku University developed tiny magnetically driven spinning screws intended to swim along veins and carry drugs to infected tissues or even to burrow into tumors and kill them with heat. In 2003, the "MR-Sub" project of Martel's group at the Nano Robotics Laboratory of Poly technique in Montreal tested using variable MRI magnetic fields to generate forces on an un lathered micro robot containing ferromagnetic particles, developing sufficient propulsive power to direct the small device through the human body. Brad Nelson's team at the Swiss Federal Institute of Technology in Zurich continued this approach. In 2005, they reported the fabrication of a microscopic robot small enough (~200 μm) to be injected into the body through a syringe. They hope that this device or its descendants might someday be used to deliver drugs or perform minimally invasive eye surgery. Nelson's simple micro robot has successfully maneuvered through a watery maze using external energy from magnetic fields, with different frequencies that are able to vibrate different mechanical parts on the device to maintain selective control of different functions.

### 4.2 Manufacturing medical nanorobots

The greatest power of nanomedicine will emerge, perhaps in the 2020s, when we can design and construct complete artificial nanorobots using rigid diamondoid nanometer-scale parts like molecular gears and bearings. These nanorobots will possess a full panoply of autonomous subsystems including onboard sensors, motors, manipulators, power supplies, and molecular computers.

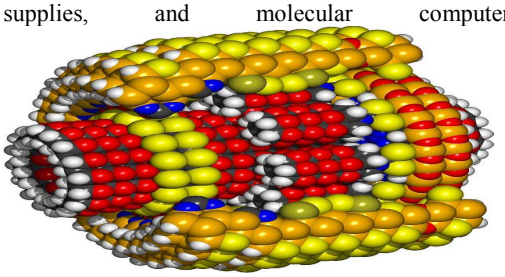


**Figure 1** A molecular planetary gear is a mechanical component that might be found inside a medical nanorobot. The gear converts shaft power from one angular frequency to another. The casing is a strained silicon shell with predominantly sulfur termination, with each of the nine planet gears attached to the planet carrier by a carbon carbon single bond. The planetary gear shown here has not been built experimentally but has been modeled computationally. .

The positional assembly of diamondoid structures, some almost atom by atom, using molecular feedstock has been examined theoretically via computational models of diamond mechanosynthesis (DMS). DMS is the controlled addition of carbon atoms to the growth surface of a diamond crystal lattice in a vacuum-manufacturing environment.
Covalent chemical bonds are formed one by one as the result of positionally constrained mechanical forces applied at the tip of a scanning probe microscope apparatus, following a programmed sequence. Mechanosynthesis using silicon atoms was first achieved experimentally in 2003. Carbon atoms should not be far behind.

For example, simple mechanical ciliary arrays consisting of 10,000 independent microactuators on a 1-$cm^2$ chip have been made at the Cornell National Nanofabrication Laboratory for microscale parts transport applications, and similarly at IBM for mechanical data storage applications.

### 4.3 Respirocytes and microbivores

The ability to build complex diamondoid medical

nanorobots to molecular precision, and then to build them cheaply enough in sufficiently large numbers to be useful therapeutically, will revolutionize the practice of medicine and surgery. The first theoretical de sign study of a complete medical nanorobot ever published in a peer-reviewed journal (in 1998) described a hypothetical artificial mechanical red blood cell or ‘‘respirocyte’’ made of 18 billion precisely arranged structural atoms. The respirocyte is a bloodborne spherical 1- µm diamondoid 1000-atmosphere pressure vessel with reversible molecule-selective surface pumps powered by endogenous serum glucose.

The nanorobots do not increase the risk of sepsis or septic shock because the pathogens are completely digested into harmless sugars, amino acids and the like, which are the only effluents from the nanorobot.

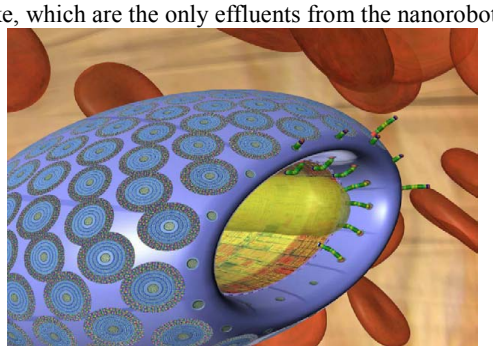


**Figure 2** Nanorobotic artificial phagocytes called ‘‘microbivores’’ could patrol the bloodstream, seeking out and digesting unwanted pathogens.

## 5. Conclusion

Data warehouses using a *multidimensional* view of data are increasingly used for business as well as scientific purposes such as medicine and bio-chemistry. However, the application of DW technology to the scientific domain poses several great challenges to existing DW technology. This paper presented techniques that aimed to solve two such challenges, namely the optimal use of precomputed aggregates over irregular hierarchies and the integration of multidimensional data with complex external data.

The first technique employed the process of *normalizing* irregular dimension hierarchies in order to enable practical pre-aggregation. When the dimension hierarchies are irregular, we showed that this technique is far superior to the two alternatives, namely *no* or *full* pre-aggregation.

The second technique used a *federation* approach to logically combine multidimensional OLAP data with complex external data stored in object databases. The paper showed that the federation approach was often superior, in terms of flexibility, correct query results, and performance, to physical integration of the data.

## 6. References


[1] Microsoft Corporation. OLE DB for OLAP Version 1.0 Specification. Microsoft Technical Document, 1998.

[2] The OLAP Report. *Database Explosion*. <www.olapreport.com/DatabaseExplosion.htm> . Current as of February 18, 2000.

[3] T. B. Pedersen and C. S. Jensen. Research Issues in Clinical Data Warehousing. In *Proceedings of the Tenth International Conference on Statistical and Scientific Database Management*, pp. 43–52, 1998.

[4] T. B. Pedersen, C. S. Jensen, and C. E. Dyreson. Supporting Imprecision in Multidimensional Databases Using Granularities. In *Proceedings of the Eleventh International Conference on Statistical and Scientific Database Management*, pp. 90–101, 1999.

[5] T. B. Pedersen, C. S. Jensen, and C. E. Dyreson. Extending PractiPre-Aggregation in On-Line Analytical Processing. In *Proceedings of the Twentyfifth International Conference on Very Large Data Bases*, pp. 663–674, 1999.

[6] T. B. Pedersen and C. S. Jensen. Multidimensional Data Modeling for Complex Data. In *Proceedings of the Fifteenth International Conference on Data Engineering*, 1999. Extended version available as TimeCenter Technical Report TR-37, <www.cs.auc.dk/TimeCenter>, 1998.