B.E.

Seventh Semester Examination, December-2008

Data Warehousing and Data Mining (IT-401-E)

Note : Attempt any *five* questions. All questions carry equal marks.

Q. 1. (a) Differentiate the following :

(i) Operational Databases & Data Warehouse

(ii) ROLAP and MOLAP

Ans. (i) Operational Databases & Data Warehouse :

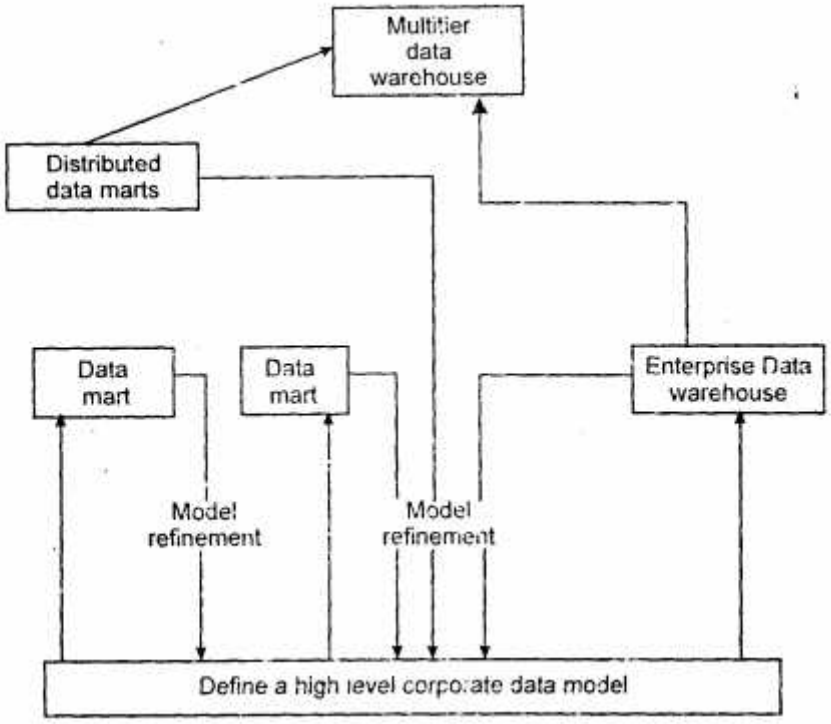
Feature	OLTP	OLAP
Characteristic	Operational processing	Information processing
Orientation	Transaction	Analysis
User	Clerk	Manager
Function	day-to-day	Long-term operations
DB design	ER Based	Subject oriented
Data	Current	Historical
Summarization	Printative	Summarized
View	Detailed	Summarized
Unit of work	Short	Complex
Access	Read	Mostly Read
Focus	Data in	Info. out
DB size	100 Mb-Gb	100 Gb to TB

(ii) ROLAP and MOLAP :

The middle tier of three tier Data warehousing architecture is an OLAP server that is typically implemented using either.

(i) relational OLAP (ROLAP) model that is an extended relational DBMS that maps operations on multi dimensional data to standard relational operations; or

(ii) a multidimensional OLAP (MOLAP) model, that is a special purpose server that directly implements multidimensional data & operations.



A recommended approach for data warehouse development

Q. 1. (b) Discuss various applications of data warehousing & Data mining.

Ans. Data Mining Applications :

(i) For Financial Data analysis

(ii) For the Retail Industry

(iii) For Telecommunication Industry

(iv) For Biological Data analysis

(v) In other scientific applications

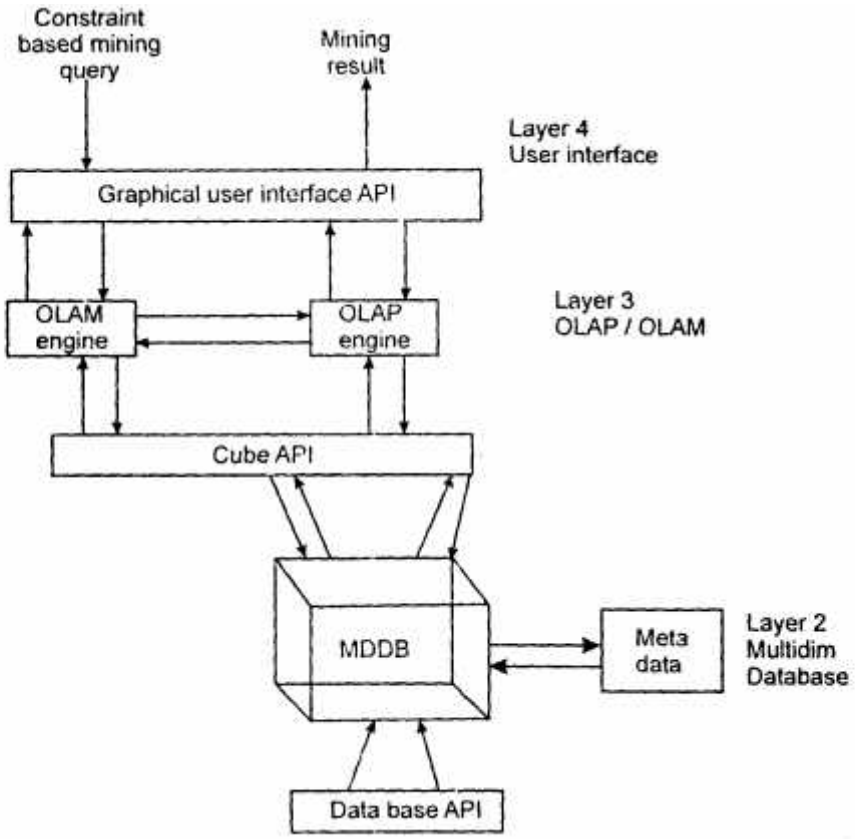
(vi) For intrusion Detection

Data Warehousing Applications :

(i) Information processing

(ii) Analytical processing

(iii) Data mining supports knowledge, Discovery for finding hidden patterns and associations.



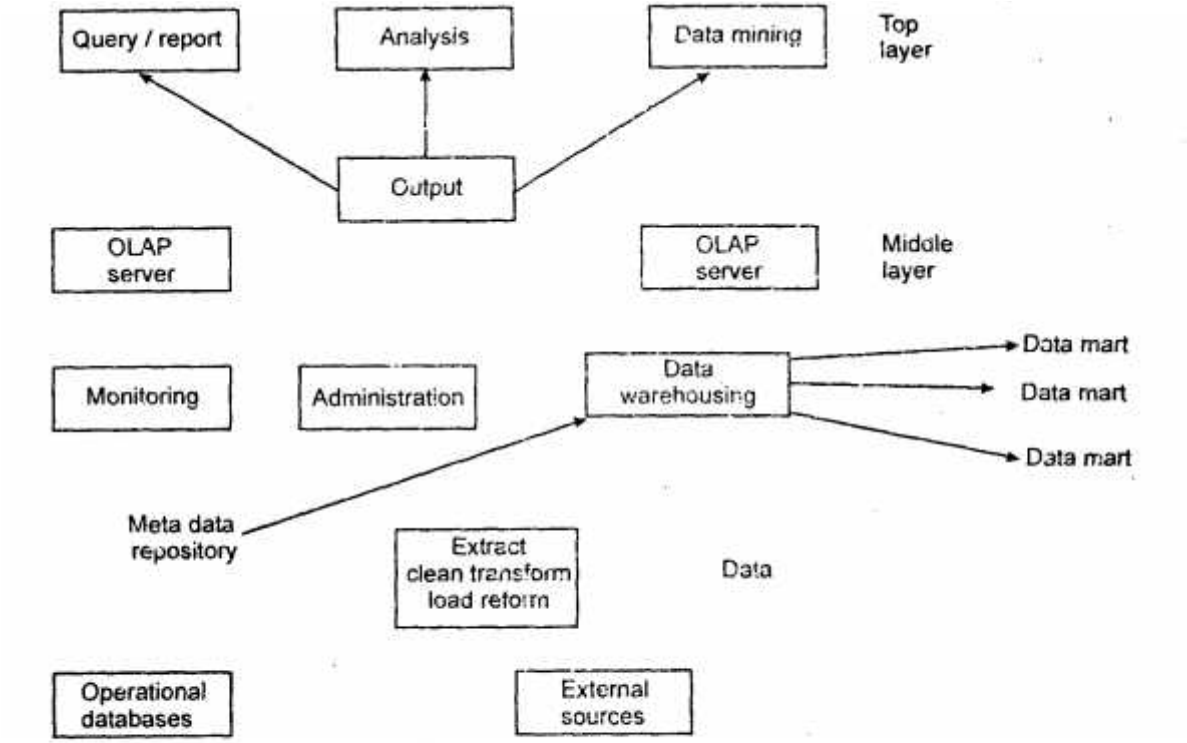
An integrated OLAM/OLAP architecture

Q. 2. (a) Explain in detail the three-tier data warehouse architecture.

Ans. ROLAP and MOLAP : The middle tier of three tier Data warehousing architecture is an OLAP server that is typically implemented using either.

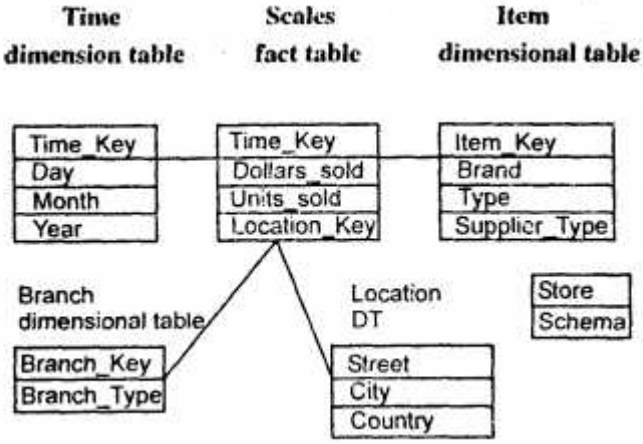
(i) relational OLAP (ROLAP) model that is an extended relational DBMS that maps operations on multi dimensional data to standard relational operations; or

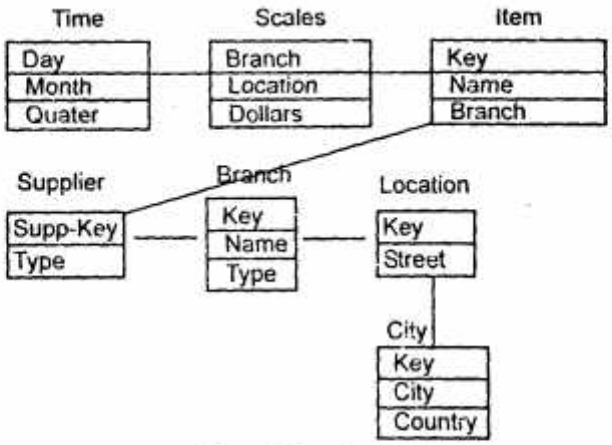
(ii) a multidimensional OLAP (MOLAP) model, that is a special purpose server that directly implements multidimensional data & operations.



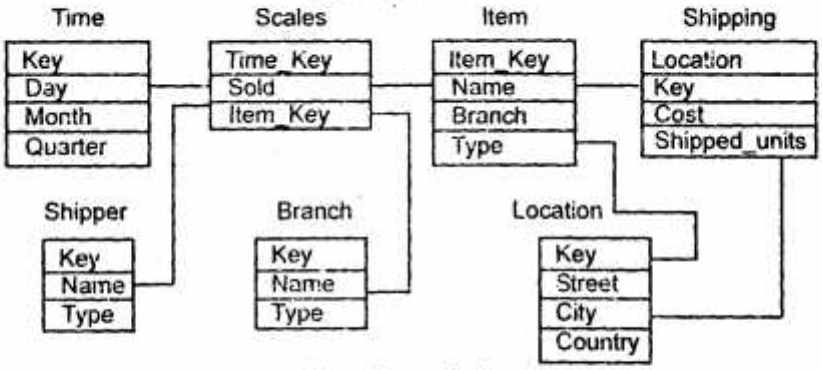
Q. 2. (b) Discuss the concept of star, snowflake & galaxy schema. How can one define the respective schemas in a data mining query language ? Take some suitable examples.

Ans.





Snowflake schema



Fact Constellation

Examples :

Cube definitions

define cube < cube name> < dimension list > < measure list >

define cube sales_star [time, branch, location] :

dollar_sold = [sum (sales)] units = count (*)

Q. 3. (a) What are the different steps involved in KDD ? What is data mining & how it is different from KDD and data warehousing ?

Ans. (i) Data Cleaning : To remove noise and inconsistent data.

(ii) Data Integration : Where multiple data sources may be combined.

(iii) Data Securian : Where data relevant to the analysis task are retrieved from the data base.

(iv) Data Transformation : Where data are transformed/consolidated into forms approximate for mining by performing summary or aggregation operations.

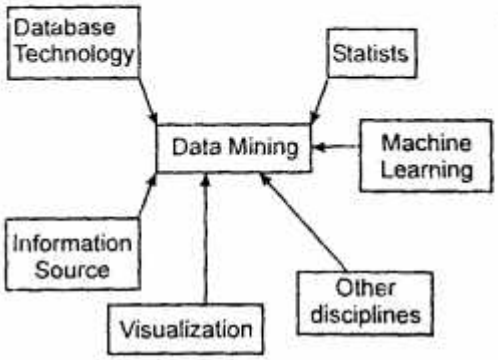
(v) Data Mining : An essential process where intelligent methods are applied in order to extract data patterns.

(vi) Pattern Evaluation : To identify the truly interesting patterns representing knowledge based on some interestingness measures.

(vii) Knowledge Presentation : Where visualization & knowledge representation techniques are used to present the mixed knowledge to user.

Q. 3. (b) What is meant by classification rules ? Explain fuzzy set, Rough set & genetic algorithm approach used in knowledge discovery through classification.

Ans.



Data Mining or a confluence of Multiple disciplines

(i) Genetic Algorithm are easily parallelizable and have been used for classification as well as other optimization problems. In data mining, they may be used to evaluate the fitness of other algorithms.

Off spring are created by applying genetic operations such as cross over and mutation. In crossover, substrings from pairs of rules are swapped to form new pairs of sum. In mutation, randomly selected bits in a rule's string are inverted.

(ii) Rough Set Appruach : Rough set theory can be used for classification to discover structural relationships within impricise or noisy data or applies to discrete valued attributes continuous valued attributes must therefore be discretised before its use.

Rough set theory is based on the establishment of the equivalence classes within the given training data.

(iii) Fuzzy Set Approaches : Rule based systems for classification have the disadvantage that they involve sharp antol is for continuous attribute. For example, consider the rule for customer credic application approval. The rule essentially says that applications for customers who have had a job for 200 more years & who have high income are approved.

Q. 4. (a) What is a priori algorithm to find frequent item sets ? Apply this algorithm to find frequent item sets in an example, Database & then find the association rules. Assume suitable measures.

Ans. A priori is a seminal algorithm proposed by R. Agrawal & R. Srikant in 1994 for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses aprior knowledge of frequent itemset properties, as we shall see following :

Apriori Property : All non-empty subsets of a frequent itemset must also be frequent.

Transaction data for All Electronics Branch

T10	List of item IDS
T100	I1, I2 , I5
T200	I2, I5
T300	I2, I3
T400	I1, I2, I3

T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Scan D for count of each candidate →

Itemset	Sup count
I1	6
I2	7
I3	6
I4	2
I5	2

→ Compare candidate support count with Minimum support count

L_1

Item set	Sup count
I1	6
I2	7
I3	6
I4	2

→ Generate C_2 candidates from L_1

C_2

Item set	Sup. count
I1, I2	4
I1, I3	4
I1, I4	1
I1, I5	2
I1, I3	4
I2, I5	2
I2, I4	2
I2, I5	2
I2, I5	0
I2, I5	1
I2, I6	0

Compare Candidate Support count with minimum Support count →

L_2

Itemset	Sup. count
I1, I2	4
I1, I3	4
I1, I5	2
I2, I3	4

Q. 4. (b) Write a short note on multimedia Databases.

Ans. Multimedia databases store image, audio & video data. They are used in applications such as picture, content based retrieval, voice mail system, video on annual systems & world wide web & speech based user interfaces that recognize spoken commands multimedia databases must support large objects because data objects such as video can require giga bytes of storage, specialized storage & search techniques are also required. Because video and audio require real time removal at a steady and predetermined rate in order to avoid picture or source gaps and system buffer overflows, such data are reffered to as continuous media data.

Q. 5. (a) Discuss various functionalities of load manager, data warehouse manager & query manager employed in Data warehousing.

Ans. (i) A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies & derived data definitions as well as data mart locations & contents.

(ii) Operational meta data, which include data image, currency of data & monitoring information.

(iii) The algorithms used for summarization, which include measure & dimensions definition algorithms data on granularity, partitions, subject areas, aggregation, summarization & predefined queries & reports.

(iv) Business metadata, which include business terms & definition data ownership information & charging policies.

(v) Data related to system performance, which indices & profiles that improve data access and retrieval performance, in addition to rules for the timing & scheduling of reform, update and replication cycles.

Q. 5. (b) What are back end tools & utilities in data warehousing ?

Ans. Data warehouse systems use back-end tools & utilities to populate refresh their data. These tools & utilities include the following functions :

(i) Data Extraction which typically gathers data from multiple, hetrogeneous & external sources.

Data leaving which detects errors in the data & rectifies them when possible.

Data transformation which converts data from legacy or host-format to warehouse format.

Load, which sets, summarizes, consolidates, computes views, checks integrity, builds indices & partitions.

Refresh, which propagates the updates from the data sources to the warehouse.

Q. 6. (a) What is meant by data mining query language ? How pattern presentation visualization specification can be done in this language ?

Ans. Data mining is an exploratory process. An easy to use and high quality graphical user interface is essential in order to promote user guided, highly interactive data mining.

Most data mining systems provide user friendly interfaces for mining. However, unlike relational databases systems, where most graphical user interfaces are constructed on top of SQL, most data mining systems donot share any underlying data mining query language. Lack of a standard data mining language makes it difficult to standardize data mining products and to ensure the interoperability of data mining system. Recent efforts at defining & standardizing data mining, which is described in the appendix.

Q. 6. (b) What is web mining, explain ? What are the various outcomes which are expected after mining WWW ?

Ans. The www serves as a huge, widely distributed, global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce & many other information services.

The web seems to be too huge for effective data warehousing & data mining, The size of the web is in the order of hundreds of tetra bytes & is ethic growing rapidly. Many organisations & societies place most of their public accessible information on the web.

The complexity of web pages is far greater than that of any traditional text document collection.

The web is a highly dynamic information source.

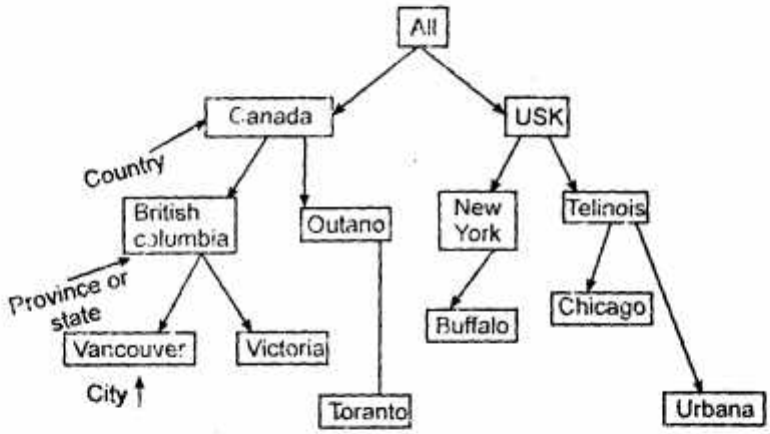
The web serves a broad diversity of user communities. Only a small portions of the information on the web is truely relevant or useful. It is said that 99% of the web information is useless to 99% of web users.

These challenges have promoted searches into effective & effective discovery & use of resources on the internet.

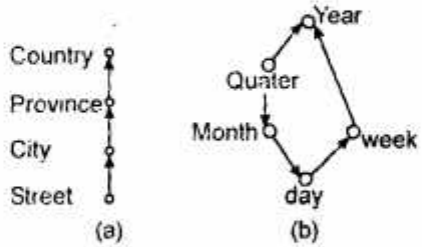
Q. 7. (a) Explain in detail the process of data warehousing design. Also state which approach is better & why ? Top down or bottom up.

Ans. A concept hierarchy defines a sequence of mappings from a set a low-level concepts to higher level, more general concepts.

Many concept hierarchies are implicit within the database schema. For example, suppose that the dimension location is described by attributes number, street, city, province, or state, zip code and country.



A concept hierarchy for the dimension location



(a) a hierarchy

(b) a lattice

Q. 7. (b) Explain the concept of Data Dictionary & Concept Hierarchy. What are the uses of constructing them in data warehouse ?

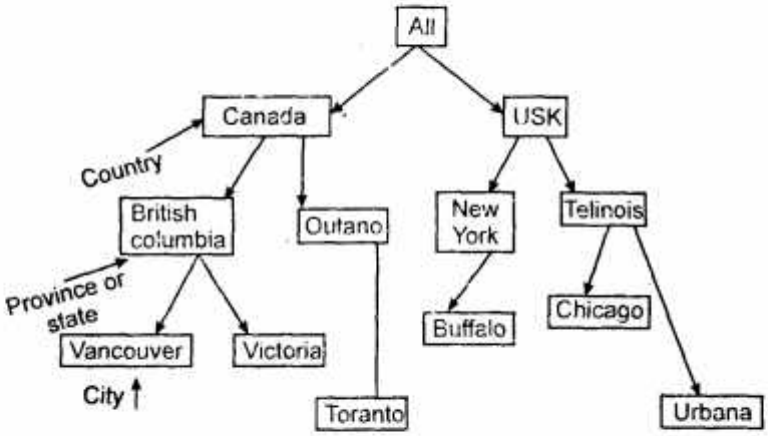
Ans. A concept hierarchy that is a total or partial order among attributes in a database schema is called a schema hierarchy. Concept hierarchies that are common to many applications may be predefined in the

mining system, such as the concept hierarchy for time. Data mining systems should provide users with the flexibility & tailor predefined hierarchy according to their practical needs.

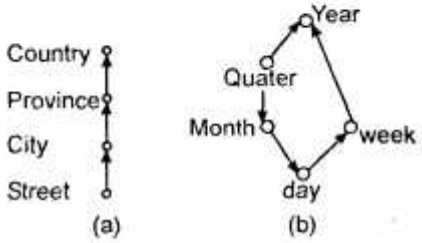
For example, users may like to define a fiscal year starting on April 1 or an academic year starting from September 1.

A concept hierarchy defines a sequence of mappings from a set a low-level concepts to higher level, more general concepts.

Many concep: hierarchies are implicit within the database schema. For example, suppose that the dimension location is described by attributes number, street, city, province, or state, zip code and country.



A concept hierarchy for the dimension location.



(a) a hierarchy
(b) a lattice

Q. 8. Write short notes on :

(a) Tuning and testing of data warehouse

(b) Support Vector Machines

(c) Complex aggregation at multiple granularities

Ans. (a) Tuning and Testing of Data Warehouse :

Data Mining Applications :

(i) For Financial Data analysis
(ii) For the Retail Industry
(iii) For Telecommunication Industry
(iv) For Biological Data analysis

(v) In other scientific applications

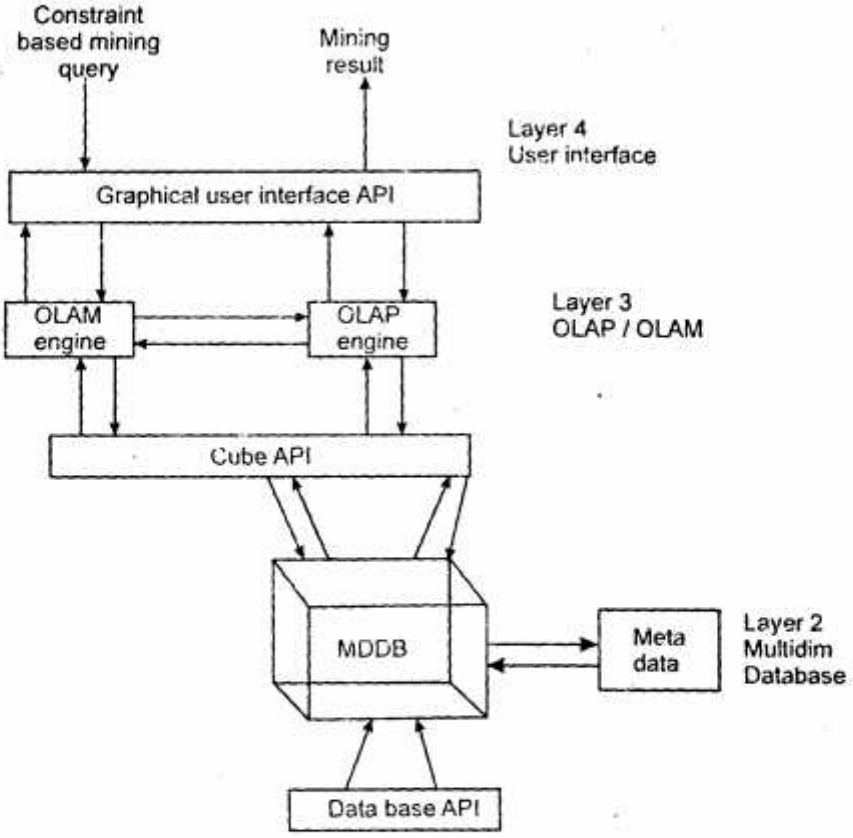
(vi) For intrusion Detection

Data Warehousing Applications :

(i) Information processing

(ii) Analytical processing

(iii) Data mining supports knowledge, Discovery for finding hidden patterns and associations.



An integrated OLAM/OLAP architecture

(b) Support Vector Machines : It is a promising new method for the classification of both linear & non-linear data. In a nutshell, a support, vector machine is an algorithm that works as follows :

It uses a non-linear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane.

Although the train time of even the fasters SVMs can be extremely slow, they are highly accurate, owing to their ability to model complex non-linear decisions boundaries. They are much less prove to overfitting that other methods. The support vectors found also province a compact description of the learned model SVMs can be used for prediction as well as classification.

They have been applied to a number of areas, including hand written digit recognition, object recognition & speaker identification, as well as benchmark time series prediction tests.