DIGITIMES

Editorial

We would like to wholeheartedly thank our honorable Chairman, **Lion.Dr.K.S.Rangasamy** and vice chairman **Mr.R.Srinivasan**, and Principal **Dr.M.Venkatesan** for their continuous encouragement and constant support for bringing out the magazine. We profoundly thank our Head of the Department **Dr.P.MeenakshiDevi** for encouraging and motivating us to lead the magazine a successful one right from the beginning. **DIGITIMES** serves as a platform for updating and enhancing upcoming technologies in Information Technology. We are also grateful to all the contributors and faculty coordinator to bring this magazine.

By,

Editorial Board

CONTENTS

DATA MINING

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science with an overall goal to extract information (with intelligent method) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

By,

HARI PRASATH K II/IT

WHY MINE DATA?

Lots of data is being collected and warehoused

- Web data, e-commerce purchases
- at department/ grocery stores
- Bank/Credit Card transactions

Computers have cheaper and more powerful. Competitive Pressure is Strong towards providing better, customized services for an edge (e.g. in Customer Relationship Management)

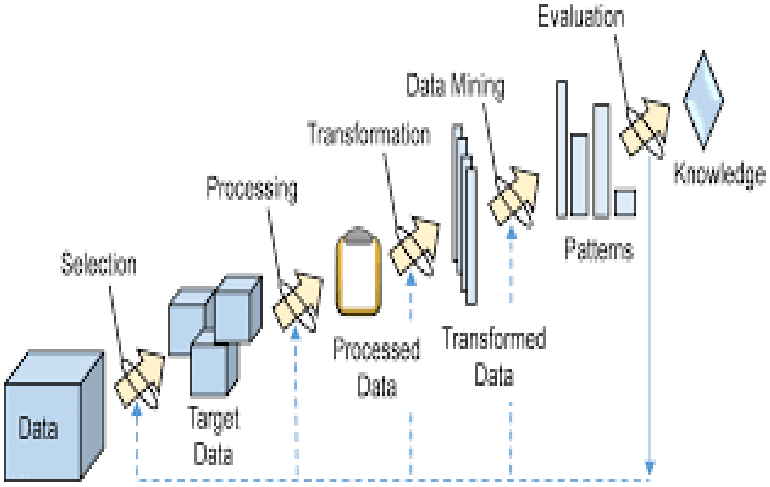
Data explosion problem: Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories. We are drowning in data, but starving for knowledge

Solution: data mining : Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases.

By,

JOSHUA DANIEL B II/IT

DATA MINING STEPS



1. **Data Integration:**

First of all the data are collected and integrated from all the different sources.

2. **Data Selection:**

In this step select only those data which is useful for data mining.

3. **Data Cleaning:**

The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies.

4. **Data Transformation:**

The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.

5. **Data Mining:**

Now we are ready to apply data mining techniques on the data to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.

6. **Pattern Evaluation and Knowledge Presentation:**

This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.

7. Decisions / Use of Discovered Knowledge:

This step helps user to make use of the knowledge acquired to take better decisions.

By,

PRIYADHARSHINI P II/IT

DATA MINING TECHNIQUES

Data Mining Techniques

The art of data mining has been constantly evolving. There are a number of innovative and intuitive techniques that have emerged that fine-tune data mining concepts in a bid to give companies more comprehensive insight into their own data with useful future trends. Many techniques are employed by the data mining experts, some of which are listed below:

Seeking Out Incomplete Data

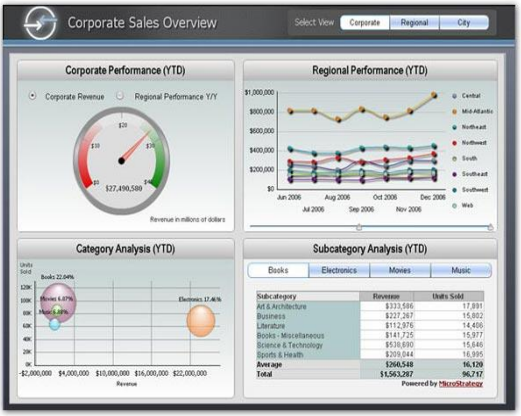
Data mining relies on the actual data present, hence if data is incomplete, the results would be completely off-mark. Hence, it is imperative to have the intelligence to sniff out incomplete data if possible. Techniques such as Self-Organizing-Maps (SOM's), help to map missing data based by visualizing the model of multi-dimensional complex data.

Multi-task learning for missing inputs, in which one existing and valid data set along with its procedures is compared with another

compatible but incomplete data set is one way to seek out such data. Multi-dimensional preceptors using intelligent algorithms to build imputation techniques can address incomplete attributes of data.

Dynamic Data Dashboards:

This is a scoreboard, on a manager or supervisor's computer, fed with real-time from data as it flows in and out of various databases within the company's environment. Data mining techniques are applied to give live insight and monitoring of data to the stakeholders.



Database Analysis:

Databases hold key data in a structured format, so algorithms built using their own language (such as SQL macros) to find hidden patterns within organized data is most useful. These algorithms are sometimes inbuilt into the data flows, e.g. tightly coupled with user-defined functions, and the findings presented in a ready-to-refer-to report with meaningful analysis.

A good technique is to have the snapshot dump of data from a large database in a cache file at any time and then analyze it further. Similarly, data mining algorithms must be able to pull out data from multiple, heterogeneous databases and predict changing trends.

Text Analysis:

This concept is very helpful to automatically find patterns within the text embedded in hordes of text files, word-processed files, PDFs, and presentation files. The text-processing algorithms can for instance, find out repeated extracts of data, which is quite useful in the publishing business or universities for tracing plagiarism.

Efficient Handling of Complex and Relational Data:

A data warehouse or large data stores must be supported with interactive and query-based data mining for all sorts of data mining functions such as classification, clustering, association, prediction. OLAP (Online Analytical Processing) is one such useful methodology. Other concepts that facilitate interactive data mining

are analyzing graphs, aggregate querying, image classification, meta-rule guided mining, swap randomization, and multidimensional statistical analysis.

Relevance and Scalability of Chosen Data Mining Algorithms:

While selecting or choosing data mining algorithms, it is imperative that enterprises keep in mind the business relevance of the predictions and the scalability to reduce costs in future. Multiple algorithms should be able to be executed in parallel for time efficiency, independently and without interfering with the transnational business applications, especially time-critical ones. There should be support to include SVMs on larger scale.

By,

SHIVANI N II/IT

The function of education is to teach one to think intensively and to think critically. Intelligence plus character - that is the goal of true education.

-Martin Luther King, Jr

POPULAR TOOLS FOR DATA

There are many readymade tools available for data mining in the market today. Some of these have common functionalities packaged within, with provisions to add-on functionality by supporting building of business-specific analysis and intelligence.

Listed below are some of the popular multi-purpose data mining tools that are leading the trends:

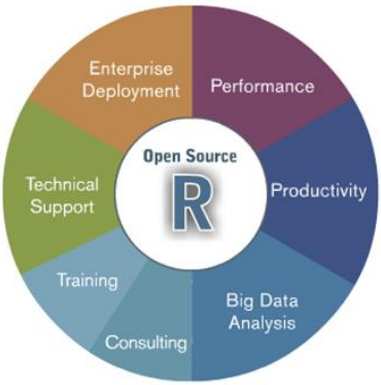
Rapid Miner (erstwhile YALE):

This is very popular since it is a ready made, open source, no-coding required software, which gives advanced analytics. Written in Java, it incorporates multifaceted data mining functions such as data preprocessing, visualization, predictive analysis, and can be easily integrated with WEKA and R-tool to directly give models from scripts written in the former two.

WEKA:

This is a JAVA based customization tool, which is free to use. It includes visualization and predictive analysis and modeling techniques, clustering, association, regression and classification.

R-Programming Tool:



This is written in C and FORTRAN, and allows the data miners to write scripts just like a programming language/platform. Hence, it is used to make statistical and analytical software for data mining. It supports graphical analysis, both linear and nonlinear modeling, classification, clustering and time-based data analysis.

Python based Orange and NTLK:

Python is very popular due to ease of use and its powerful features. Orange is an open source tool that is written in Python with useful data analytic s, text analysis, and machine-learning features embedded in a visual programming interface. NTLK, also composed in Python, is a powerful language processing data mining tool, which consists of data mining, machine learning, and data scraping features that can easily be built up for customized needs.

Knime:

Primarily used for data preprocessing – i.e. data extraction, transformation and loading, Knime is a powerful tool with GUI that shows the network of data nodes. Popular amongst financial data analysts, it has modular data pipe lining, leveraging machine learning, and data mining concepts liberally for building business intelligence reports.

Data mining tools and techniques are now more important than ever for all businesses, big or small, if they would like to leverage their existing data stores to make business decisions that will give them a competitive edge. Such actions based on data evidence and advanced analytics have better chances of increasing sales and facilitating growth. Adopting well-established techniques and tools and availing the help of data mining experts shall assist companies to utilize relevant and powerful data mining concepts to their fullest potential.

By,

PAVITHRA S II/IT

BENEFITS OF DATA MINING

- Automated prediction of trends and behaviors
- It can be implemented on new systems as well as existing platforms
- It can analyze huge database in minutes
- Automated discovery of hidden patterns
- There are a lot of models available to understand complex data easily
- It is of high speed which makes it easy for the users to analyze huge amount of data in less time
- It yields improved predictions

By,

ABHITH JOHN ANTO K IV/IT

Education is the most powerful weapon which you can use to change the world.

-Nelson Mandela

DATAMINING CONS

Privacy Issues

The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time, the personal information they own probably is sold to other or leak.

Security issues

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from the big corporation such as Ford Motor Credit Company, Sony… with so much personal

and financial information available, the credit card stolen and identity theft become a big problem.

Misuse of information/inaccurate information

Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.

By,

AISWARYA R IV/IT

Attitude is more important than the past, than education, than money, than circumstances, than what people do or say. It is more important than appearance, giftedness, or skill.

-Charles R. Swindoll

DATA MINING APPLICATIONS

Data mining in Telecommunication

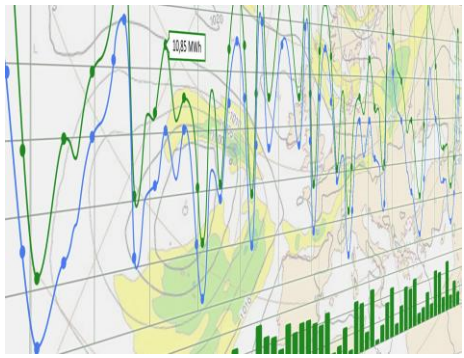
In this, data mining gains a competitive advantage and reduce customer churn by understanding demographic characteristics and predicting customer behavior.

Increases customer loyalty and improve profitability by providing customized services.

It supports customer strategy by developing appropriate marketing campaigns and pricing strategies.

Data mining for Energy

As data mining capture weak signals of potentially threatening events. Also, identify previously unidentified patterns, connections.



Structure identification of important information, and distill it to boost technical problem-solving. Also, empower more informed

decision-making and enable immediate notification of prospective technical breakthroughs.

Improve core processes in upstream, midstream and downstream. As with analysis and intelligence capabilities using a variety of sources.

Data Mining Applications in Sales and Marketing

Basically, it enables businesses to understand the hidden patterns inside historical purchasing transaction data. Thus, helping in planning and launching new marketing campaigns.

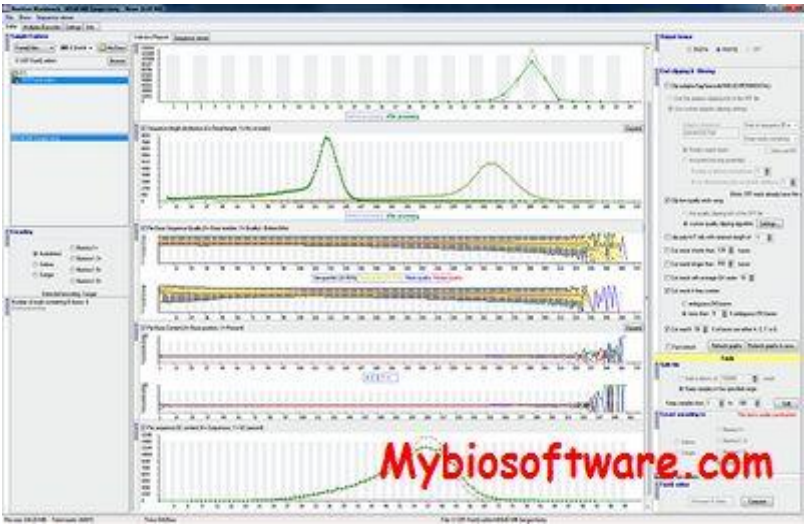
Generally, the following illustrates several data mining applications in sale and marketing.

We use it for market basket analysis. That is to provide information on what product combinations have to purchased together. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products.

Retail companies use data mining to identify customer's behavior buying patterns.

Data mining in E-commerce

Many E-commerce companies are using data mining business Intelligence to offer cross- sells through their websites. One of the most famous of these is, of course, Amazon. They use sophisticated mining techniques to drive their 'People who viewed that product. Also liked this' functionality.



Data Mining in Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biologist. Such as genomics, proteomics, functional Genomics and biomedical research. Also, Biological data mining is a very important part of Bioinformatics.

Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Data Mining for Crime Agencies

Beyond corporate applications, crime prevention agencies use analytics. And Data Mining to spot trends across myriads of data. That should help with everything from where to deploy police manpower. And Particularly who to search at a border crossing. And even which intelligence to take seriously in counter-terrorism activities.

Data Mining in Retail

Generally, retailers segment customers into 'Regency, Frequency, Monetary' groups. Also, in target marketing and promotions to those different groups. A customer who spends little but often and last did so recently will be handled by a customer. Particularly who spent big but only once, and also some time ago.

Large retailers like Wal-Mart utilize information on store footfall, advertising campaign even weather forecast to predict sales and stock up accordingly.

Credit card companies mine transaction records for fraudulent use of their cards. That was based on purchase patterns of consumers. As they can deny access if your purchase patterns change drastically!

Generally, the Human Genome Project mounts up piles of data. Although, getting the data to work for humankind need to develop a new drug and weed out diseases. That will require pattern recognition in the data which is handled in bioinformatics.

As scientists use microarray data to look at the gene expressions. And also sophisticated data analysis techniques. That is employed to account for the background noise and normalization of data.

By,

BALAJI S IV/IT

Education is not preparation for life; education is life itself.

-John Dewey

DATAWAREHOUSING

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources.

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making.

How Data warehouse works?

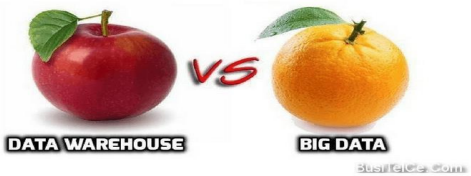
A Data Warehouse works as a central repository where information arrives from one or more data sources. Data flows into a data warehouse from the transactional system and other relational databases.

Data may be:

- Structured
- Semi-structured
- Unstructured data

The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. A data warehouse merges information coming from different sources into one comprehensive database.

By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available. Data warehousing makes data mining possible. Data mining is looking for patterns in the data that may lead to higher sales and profits.



A data warehouse is a database of a different kind: an OLAP (online analytical processing) database. A data warehouse exists as a layer on top of another database or databases . A data warehouse, on the other hand, is structured to make analytics fast and easy.

By,

DHANUSYADEVI S IV/IT

TYPES OF DATA WAREHOUSE

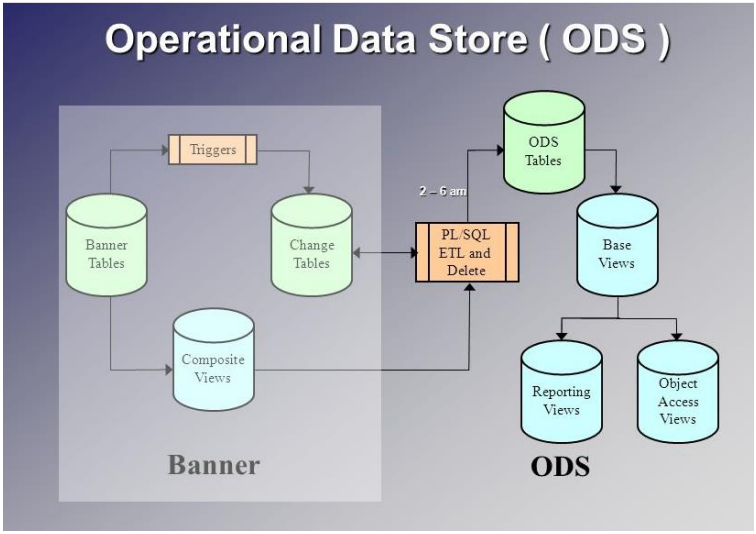
Three main types of Data Warehouses are:

1.Enterprise Data Warehouse:

Enterprise Data Warehouse is a centralized warehouse. It provides decision support service across the enterprise. It offers a unified approach for organizing and representing data. It also provide the ability to classify data according to the subject and give access according to those divisions.

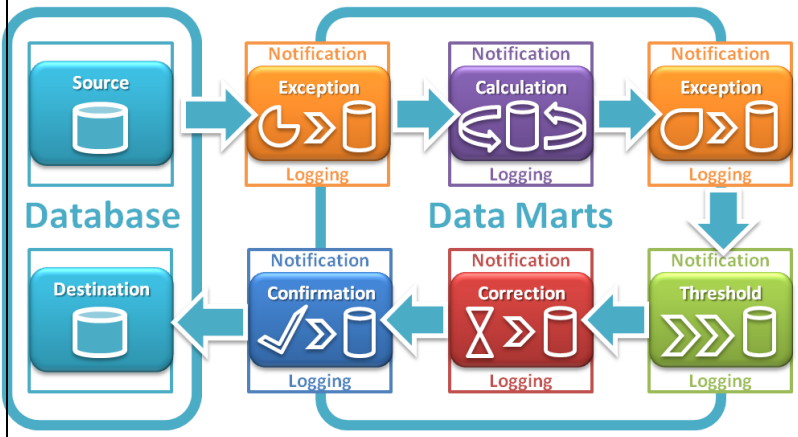
2. Operational Data Store:

Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.



3. Data Mart:

A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.



By,

ELANGOVAN P IV/IT

Education is not preparation for life; education is life itself.

-John Dewey

GENERAL STAGES OF DATA

Earlier, organizations started relatively simple use of data warehousing. However, over time, more sophisticated use of data warehousing begun.

The following are general stages of use of the data warehouse:

Offline Operational Database:

In this stage, data is just copied from an operational system to another server. In this way, loading, processing, and reporting of the copied data do not impact the operational system's performance.

Offline Data Warehouse:

Data in the Data warehouse is regularly updated from the Operational Database. The data in Data warehouse is mapped and transformed to meet the Data warehouse objectives.

Real time Data Warehouse:

In this stage, Data warehouses are updated whenever any transaction takes place in operational database. For example, Airline or railway booking system.

Integrated Data Warehouse:

In this stage, Data Warehouses are updated continuously when the operational system performs a transaction. The Data warehouse then generates transactions which are passed back to the operational system.

By,

MOHANA PRIYA S IV/IT

COMPONENTS OF DATA

Four components of Data Warehouses are:

Load manager:

Load manager is also called the front component. It performs with all the operations associated with the extraction and load of data into the warehouse. These operations include transformations to prepare the data for entering into the Data warehouse.

Warehouse Manager:

Warehouse manager performs operations associated with the management of the data in the warehouse. It performs operations like analysis of data to ensure consistency, creation of indexes and views, generation of denormalization and aggregations, transformation and merging of source data and archiving and baking-up data.

Query Manager:

Query manager is also known as backend component. It performs all the operation operations related to the management of user queries. The operations of this Data warehouse components are direct queries to the appropriate tables for scheduling the execution of queries.

End-user access tools:

This is categorized into five different groups like

1. Data Reporting
2. Query Tools
3. Application development tools
4. EIS tools
5. OLAP tools and data mining tools.

By,

BOOPALAN G III/IT

STEPS TO IMPLEMENT DATA

Steps to Implement Data Warehouse

The best way to address the business risk associated with a Data warehouse implementation is to employ a three-prong strategy as below

Enterprise strategy: Here we identify technical including current architecture and tools. We also identify facts, dimensions, and attributes. Data mapping and transformation is also passed.

Phased delivery: Data warehouse implementation should be phased based on subject areas. Related business entities like booking and billing should be first implemented and then integrated with each other.

Iterative Prototyping: Rather than a big bang approach to implementation, the Data warehouse should be developed and tested iteratively.

By,

HAKKEM M III/IT

ADVANTAGES OF DATA WAREHOUSE

1. Integrating data from multiple sources;

2. Performing new types of analyses

3. Reducing cost to access historical data.

4. Improving turnaround time for analysis and reporting;

5. Sharing data and allowing others to easily access data;

6. Supporting ad hoc reporting and inquiry;

7. Reducing the development burden on IS/IT; and

8. Removing informational processing load from transaction-oriented databases;

By,

MOHANRAJ S III/IT

DISADVANTAGES OF DATA

Data warehouses are relational databases that act as data analysis tools, aggregating data from multiple departments of a business into one data store. Data warehouses are typically updated as an end-of-day batch job, rather than being churned by real time transactional data. Their primary benefits are giving managers better and timelier data to make strategic decisions for the company. However, they have some drawbacks as well.

Extra Reporting Work

Depending on the size of the organization, a data warehouse runs the risk of extra work on departments. Each type of data that's needed in the warehouse typically has to be generated by the IT teams in each division of the business. This can be as simple as duplicating data

from an existing database, but at other times, it involves gathering data from customers or employees that wasn't gathered before.

Cost/Benefit Ratio

A commonly cited disadvantage of data warehousing is the cost/benefit analysis. A data warehouse is a big IT project, and like many big IT projects, it can suck a lot of IT man hours and budgetary money to generate a tool that doesn't get used often enough to justify the implementation expense. This is completely sidestepping the issue of the expense of maintaining the data warehouse and updating it as the business grows and adapts to the market.

Data Ownership Concerns

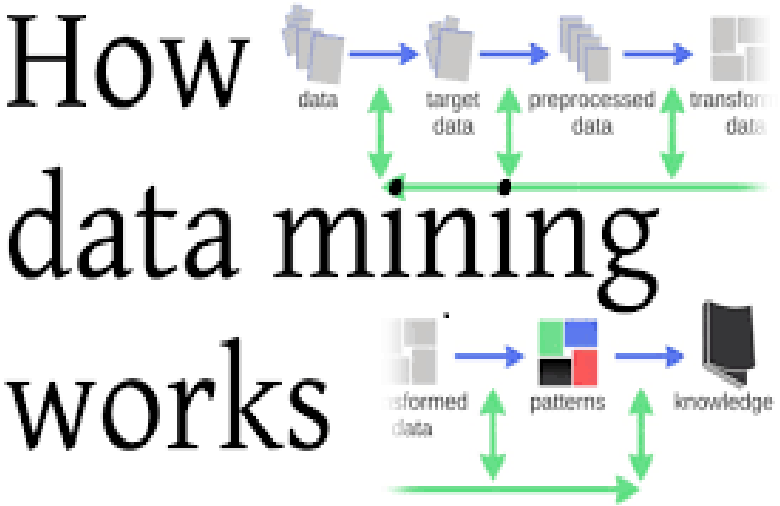
Data warehouses are often, but not always, Software as a Service implementations, or cloud services applications. Your data security in this environment is only as good as your cloud vendor. Even if implemented locally, there are concerns about data access throughout the company. Make sure that the people doing the analysis are individuals that your organization trusts, especially with customers' personal data. A data warehouse that leaks customer data is a privacy and public relations nightmare.

Data Flexibility

Data warehouses tend to have static data sets with minimal ability to "drill down" to specific solutions. The data is imported and filtered through a schema, and it is often days or weeks old by the time it's actually used. In addition, data warehouses are usually subject to ad hoc queries and are thus notoriously difficult to tune for processing speed and query speed. While the queries are often ad hoc, the queries are limited by what data relations were set when the aggregation was assembled.

By,

NESHAK KUMAR S III/IT



DATA WAREHOUSE APPLICATIONS

Banking Industry

In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.

Most banks also use warehouses to manage the resources available on deck in an effective manner. Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.

Analysis of card holder's transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity. Apart from all these, there is also scope for co-branding.

Finance Industry

Similar to the applications seen in banking, mainly revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

Consumer Goods Industry

They are used for prediction of consumer trends, inventory management, market and advertising research. In-depth analysis of sales and production is also carried out. Apart from these, information is exchanged business partners and clientele.

Government and Education

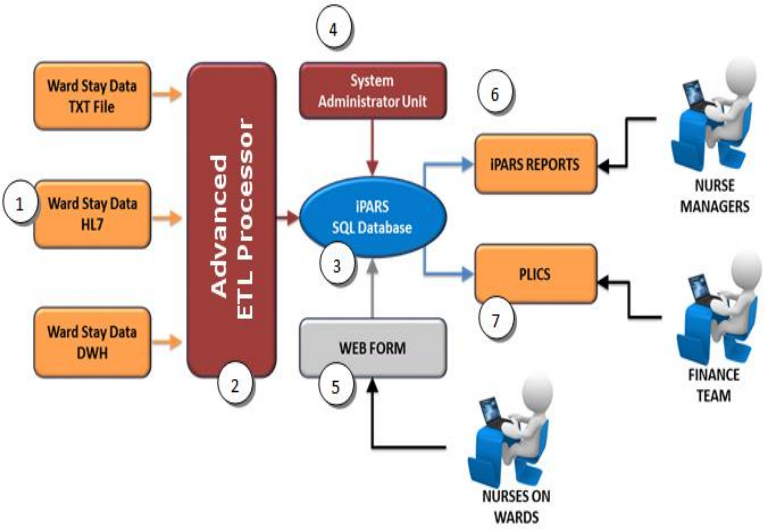
The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.

The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers, and also their entire criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.

Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management. The entire financial department of most universities depends on data warehouses, inclusive of the Financial Aid department.

Healthcare

One of the most important sector which utilizes data warehouses is the Healthcare sector. All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate



patient reports, share data with tie-in insurance companies, medical aid services, etc.

Hospitality Industry

A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services. They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.

Insurance

As the saying goes in the insurance services sector, "Insurance can never be bought, it can be only be sold", the warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants. The design of tailor-made customer offers and promotions is also possible through warehouses.

By,

SAKTHI RAVI KUMAR M III/IT

.

META DATA

- Metadata is simply defined as **data about data.**
- The data that is **used to represent other data** is known as metadata.
- For example, the **index of a book** serves as a metadata for the contents in the book.
- In other words, we can say that metadata is the **summarized data** that leads us to detailed data.

1. In terms of data warehouse

- Metadata is the **road-map** to a data warehouse.
- Metadata in a data warehouse **defines the warehouse objects**.
- Metadata acts as a **directory**. This directory **helps the decision support system** to locate the contents of a data warehouse.

2.Categories of Metadata

Metadata can be broadly categorized into three categories:

- **Business Metadata**
  - It has the data ownership information, business definition, and changing policies.
- **Technical Metadata**

- It includes database system names, table and column names and sizes, data types and allowed values.
- Technical metadata also includes structural information such as primary and foreign key attributes and indices.

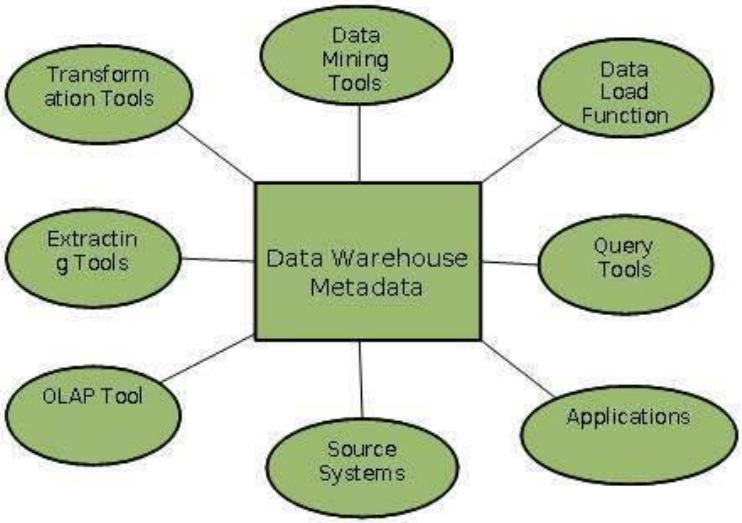
- **Operational Metadata**
  - It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged.

3. Role of Metadata

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for **mapping of data when data is transformed from operational environment** to data warehouse environment.
- Metadata is used for query tools.
- Metadata is used in extraction, cleansing, reporting transformation tools.
- Metadata plays an important role in loading functions.

4. Metadata Repository

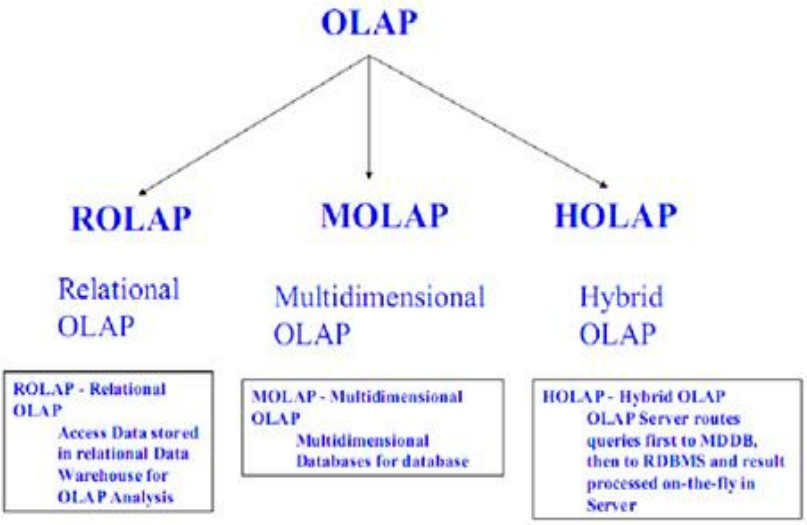
- Metadata repository is an **integral part of a data warehouse system**.
- It has the following metadata:
  - **Definition of data warehouse** - It includes the description of structure of data warehouse. The description is defined by schema, view, hierarchies, derived data definitions, and data mart locations and contents.
  - **Algorithms for summarization** - It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

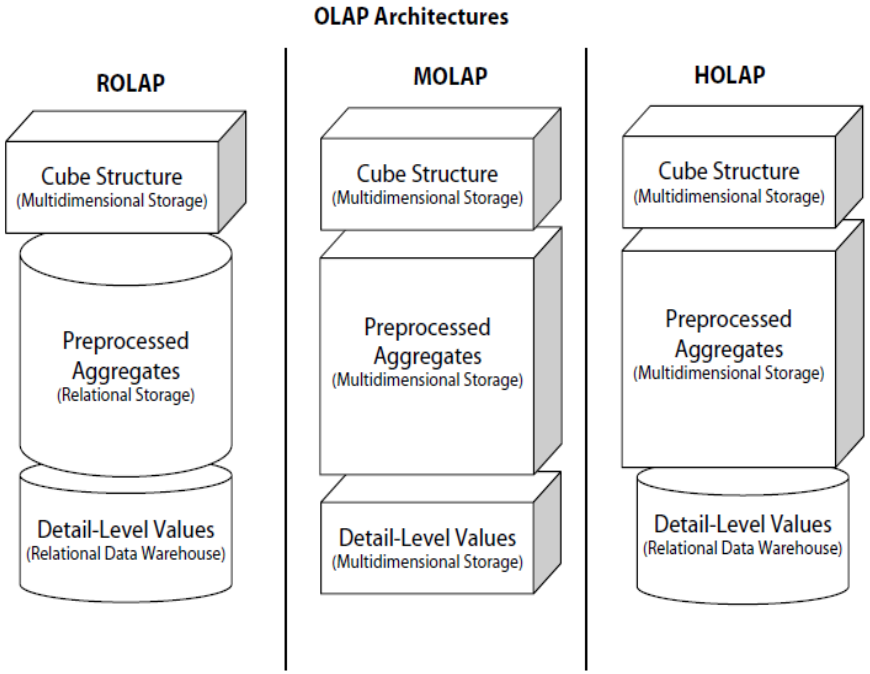


By,

ARAVIND T III/IT

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)


OLAP
ROLAP
MOLAP
HOLAP
Relational OLAP
Multidimensional OLAP
Hybrid OLAP
ROLAP - Relational OLAP
Access Data stored in relational Data Warehouse for OLAP Analysis
MOLAP - Multidimensional OLAP
Multidimensional Databases for database
HOLAP - Hybrid OLAP
OLAP Server routes queries first to MDDB, then to RDBMS and result processed on-the-fly in Server


OLAP Architectures
ROLAP
Cube Structure
(Multidimensional Storage)
Preprocessed Aggregates
(Relational Storage)
Detail-Level Values
(Relational Data Warehouse)
MOLAP
Cube Structure
(Multidimensional Storage)
Preprocessed Aggregates
(Multidimensional Storage)
Detail-Level Values
(Multidimensional Storage)
HOLAP
Cube Structure
(Multidimensional Storage)
Preprocessed Aggregates
(Multidimensional Storage)
Detail-Level Values
(Relational Data Warehouse)