Data Warehouse – Data Mart – Data Mining

A data warehouse is a place where data is stored for archival, analysis and security purposes. Usually a data warehouse is either a single computer or many computers (servers) tied together to create one giant computer system.

Data can consist of raw data or formatted data. It can be on various types of topics including organization's sales, salaries, operational data, summaries of data including reports, copies of data, human resource data, inventory data, external data to provide simulations and analysis, etc.

Besides being a store house for large amount of data, they must possess systems in place that make it easy to access the data and use it in day to day operations. A data warehouse is sometimes said to be a major role player in a decision support system (DSS). DSS is a technique used by organizations to come up with facts, trends or relationships that can help them make effective decisions or create effective strategies to accomplish their organizational goals.

Data Warehousing Models

There are many different models of data warehouses. Online Transaction Processing, which is a data warehouse model, is built for speed and ease of use. Another type of data warehouse model is called Online Analytical processing, which is more difficult to use and adds an extra step of analysis within the data. Usually it requires more steps which slows the process down and requires much more data in order to analyze certain queries.

In addition to this model, one of the more common data warehouse models include a data warehouse that is subject oriented, time variant, non volatile and integrated. Subject oriented means that data is linked together and is organized by relationships. Time variant means that any data that is changed in the data warehouse can be tracked. Usually all changes to data are stamped with a time-date and with a before and after value, so that you can show the changes throughout a period of time. Non volatile means that the data is never deleted or erased. This is a great way to protect your most crucial data. Because this data is retained, you can continue to use it in a later analysis. Finally, the data is integrated, which means that a data warehouse uses data that is organizational wide instead of from just one department.

Besides the term data warehouse, a term that is frequently used is a data mart. Data marts are smaller and less integrated data housings. They might be just a database on human resources records or sales data on just one division.

Types of Data Warehouses

With improvements in technology, as well as innovations in using data warehousing techniques, data warehouses have changed from Offline Operational Databases to include an Online Integrated data warehouse.

Offline Operational Data Warehouses are data warehouses where data is usually copied and pasted from real time data networks into an offline system where it can be used. It is usually the simplest and less technical type of data warehouse.

Offline Data Warehouses are data warehouses that are updated frequently, daily, weekly or monthly and that data is then stored in an integrated structure, where others can access it and perform reporting.

Real Time Data Warehouses are data warehouses where it is updated each moment with the influx of new data. For instance, a Real Time Data Warehouse might incorporate data from a Point of Sales system and is updated with each sale that is made.

Integrated Data Warehouses are data warehouses that can be used for other systems to access them for operational systems. Some Integrated Data Warehouses are used by other data warehouses, allowing them to access them to process reports, as well as look up current data.

Advantages & Disadvantages

The number one reason why you should implement a data warehouse is so that employees or end users can access the data warehouse and use the data for reports, analysis and decision making. Using the data in a warehouse can help you locate trends, focus on relationships and help you understand more about the environment that your business operates in.

Data warehouses also increase the consistency of the data and allow it to be checked over and over to determine how relevant it is. Because most data warehouses are integrated, you can pull data from many different areas of your business, for instance human resources, finance, IT, accounting, etc.

While there are plenty of reasons why you should have a data warehouse, it should be noted that there are a few negatives of having a data warehouse including the fact that it is time consuming to create and to keep operating.

You might also have a problem with current systems being incompatible with your data. It is also important to consider future equipment and software upgrades; these may also need to be compatible with you data.

Finally, security might be a huge concern, especially if your data is accessible over an open network such as the internet. You do not want your data to be viewed by your competitor or worse hacked and destroyed.

Data Mart

In some data warehouse implementations, a data mart is a miniature data warehouse; in others, it is just one segment of the data warehouse. Data marts are often used to provide information to functional segments of the organization. Typical examples are data marts for the sales department, the inventory and shipping department, the finance department, upper level management, and so on. Data marts can also be used to segment data warehouse data to reflect a geographically compartmentalized business in which each region is relatively autonomous. For example, a large service organization may treat regional operating centers as individual business units, each with its own data mart that contributes to the master data warehouse.

Data marts are sometimes designed as complete individual data warehouses and contribute to the overall organization as a member of a distributed data warehouse. In other designs, data marts receive data from a master data warehouse through periodic updates, in which case the data mart functionality is often limited to presentation services for clients.

Regardless of the functionality provided by data marts, they must be designed as components of the master data warehouse so that data organization, format, and schemas are consistent throughout the data warehouse. Inconsistent table designs, update mechanisms, or dimension hierarchies can prevent data from being reused throughout the data warehouse, and they can result in inconsistent reports from the same data. For example, it is unlikely that summary reports produced from a finance department data mart that organizes the sales force by management reporting structure will agree with summary reports produced from a sales department data mart that organizes the same sales force by geographical region. It is not necessary to impose one view of data on all data marts to achieve consistency; it is usually possible to design consistent schemas and data formats that permit rich varieties of data views without sacrificing interoperability. For example, the use of a standard format and organization for time, customer, and product data does not preclude data marts from presenting information in the diverse perspectives of inventory, sales, or financial analysis.

Data marts should be designed from the perspective that they are components of the data warehouse regardless of their individual functionality or construction. This provides consistency and usability of information throughout the organization.

Microsoft® SQL Server™ 2000 tools used for a data mart may include any of the tools used for data warehouses depending on how the data mart is designed. If the data mart is created and maintained locally and participates in the organization's data warehouse as an independent contributor, its creation and maintenance will involve all the operations of a data warehouse. If the data mart is a local access point for data distributed from a central data warehouse, only a subset of the tools may be involved.

Data Mining

Data mining is usually defined as searching, analyzing and sifting through large amounts of data to find relationships, patterns, or any significant statistical correlations. With the advent of computers, large databases and the internet, it is easier than ever to collect millions, billions and even trillions of pieces of data that can then be systematically analyzed to help look for relationships and to seek solutions to difficult problems. Besides governmental uses, many marketers use data mining to find strong consumer patterns and relationships. Large organizations and educational institutions also data mine to find significant correlations that can enhance our society.

While data mining is amoral in the fact that it only looks for strong statistical correlations or relationships, it can be used for either good or not so good purposes. For instance, many government organizations depend on data mining to help them create solutions for many societal problems. Marketers use data mining to help them pin point and focus their attention on

certain segments of the market to sell to, and in some cases black hat hackers can use data mining to steal and scam thousands of people.

How does data mining work? Well the quick answer is that large amounts of data are collected. Usually most entities that perform data mining are large corporations and government agencies. They have been collecting data for decades and they have lots of data to sift through. If you are a fairly new business or individual, you can purchase certain types of data in order to mine for your own purposes. In addition, data can also be stolen from large depositories by hackers by hacking their way into a large database or simply stealing laptops that are ill protected.

If you are interested in a small case study on how data mining is collected, used and profited off of, you can look at your local supermarket. Your supermarket is usually an extremely lean and organized entity that relies on data mining to make sure that it is profitable. Usually your supermarket employs a POS (Point Of Sale) system that collects data from each item that is purchased. The POS system collects data on the item brand name, category, size, time and date of the purchase and at what price the item was purchased at. In addition, the supermarket usually has a customer rewards program, which also is input into the POS system. This information can directly link the products purchased with an individual. All this data for every purchase made for years and years is stored in a database in a computer by the supermarket.

Now that you have a database with millions upon millions of data fields and records what are you going to do with it? Well, you data mine it. Knowledge is power and with so much data you can uncover trends, statistical correlations, relationships and patterns that can help your business become more efficient, effective and streamlined.

The supermarket can now figure out which brands sell the most, what time of the day, week, month or year is the most busiest, what products do consumers buy with certain items. For instance, if a person buys white bread, what other item would they be inclined to buy? Typically we can find its peanut butter and jelly. There is so much good information that a supermarket can use just by data mining their own data that they have collected.