

COP 4710: Database Systems Spring 2006

CHAPTER 25 – Data Warehousing – Part 1

Instructor : Mark Llewellyn
markl@cs.ucf.edu
CSB 242, 823-2790
<http://www.cs.ucf.edu/courses/cop4710/spr2006>

School of Electrical Engineering and Computer Science
University of Central Florida



Introduction to Decision Support Systems

- Organizations tend to grow and prosper as they gain a better understanding of their environment. Typically, business managers must be able to track daily transactions to evaluate how the business is performing.
- By tapping into the operational database, management can develop strategies to meet organizational goals. In addition, data analysis can provide information about short-term tactical evaluations and strategies, such as: Are our sales promotions working? What market percentage are we controlling? Are we attracting new customers?
- Managers understand that the business climate is very dynamic, and this mandates their prompt reaction to change in order to remain competitive.
- The modern business climate requires that managers approach increasingly complex problems involving a rapidly growing number of internal and external variables.



Introduction to Decision Support Systems (cont.)

- It should come as no surprise that there is a growing interest in creating support systems dedicated to facilitating quick decision making in a complex environment.
- Different managerial levels require different decision support needs.
 - For example, transaction-processing systems based on operational databases, are tailored to serve the information needs of people who deal with short-term inventory, accounts payable, or purchasing.
 - Middle-level managers and on up, focus on strategic and tactical decision making. Such managers require detailed information designed to help them make complex decisions in the face of a complex data and analysis environment.
- To support middle and upper management, information systems departments have created a number of **decision support systems** (DSSs).



Introduction to Decision Support Systems (cont.)

- Decision support is a methodology (or a series of methodologies) designed to extract information from data and to use such information as a basis for decision making. A **decision support system** (DSS) is an arrangement of computerized tools used to assist managerial decision making within a business.
- A DSS usually requires extensive data “massaging” to produce the required information.
- Once constructed the DSS is used at all levels within an organization and is often tailored to focus on specific business areas or problems such as finance, insurance, healthcare, banking, sales, and manufacturing.
- The DSS is interactive and provides ad hoc query tools to retrieve data and to display data in different formats. For example a user might:
 - Compare the relative rates of productivity growth by company division over some specified period of time.
 - Define the relationship between advertising types and sales levels. This relationship might then be used for forecasting purposes.



Introduction to Decision Support Systems (cont.)

- The DSS answers queries such as those on the previous page by combining historical operational data with business models that reflect the business activities.
- A typical DSS consists of four main components: a **data store component**, a **data extraction and filtering component**, an **end-user query tool**, and an **end-user presentation tool**.
 - The data store component is the data warehouse. Data warehouses differ from conventional databases in the types of data which are stored in them. Certainly a major component of the data warehouse is the operational database, but it goes well beyond that to include many different forms of data including external data (data from outside of the company).
 - The data extraction and filtering component is used to extract and validate data pulled from both the operational database as well as external sources. DSS data differs from purely operational data in three main areas: (1) time span, (2) granularity, and (3) dimensionality. We'll look at these in more detail later.



Operational Data vs. Decision Support Data

- Operational data and DSS data serve different purposes.
 - Most operational data are stored in a relational database in highly normalized fashion. Operational data storage is optimized to support transactions that represent daily operations. Operational data is frequently updated.
- DSS data give tactical and strategic business meaning to operational data. DSS data differs from operational data in three main areas: time span, granularity, and dimensionality.
 - **Time span:** operational data represent current transactions and represent relatively short time spans. DSS data represents a longer time frame. Managers are typically not interested in a particular sale to customer X, rather they tend to focus on sales generated in the last month or last year, or last five years. They are interested in the buying patterns of a customer or group of customers. The data tends to be historic in nature. The DSS data represents company transactions up to a given point in time: yesterday, last week, last month and so on. Data analysts should be aware that the sales invoice generated two minutes ago is not likely to be found in the DSS database.



Operational Data vs. Decision Support Data (cont.)

- **Granularity** (level of aggregation): DSS data must be presented at different levels of aggregation, from highly summarized to near-atomic. Managers at different levels in the organization require data with different levels of aggregation. It is also possible that a single problem requires data with different summarization levels. For example, if a manager must analyze sales region, they must be able to access data showing the sales by region, by city within a region, by store within a city within a region, and so on.

Drilling down data refers decomposing data into finer granularity.

Rolling up data refers to aggregating data to a higher level or more coarse granularity.

- **Dimensionality**: This is probably the most distinguishing characteristic of DSS data. From the data analysts point of view, the data are always related in many different ways. For example, if we analyze product sales to a customer during a given time span, we might ask “how many widgets of type X were sold to customer Y during the last six months?” This question tends to expand quickly to include many different data slices. For instance, we might want to know how product X fared compared to product Z during the past six months, by region, state, city, store, and customer. Both time and location become part of the picture.



Operational Data vs. Decision Support Data (cont.)

- Data analysts are always interested in developing the larger picture.
- Data analysts tend to include data from many data dimensions, a multi-dimensional view of the data.
- Operational data represent transaction as they happen, in real time. DSS data are a snapshot of the operational data at some point in time. Thus, DSS data are historic, representing a time slice of the operational data.
- Operational data and DSS data also differ in terms of transaction type and transaction volume. Operational data are characterized by update transactions. DSS data are characterized by query operations. DSS data also require periodic updates to load new summary data from operational data. Transaction volume tends to be high for operational data and low for DSS data.



Operational Data vs. Decision Support Data Summary

Characteristic	Operational Data	DSS Data
Data currency	current operations – real time data	historic data, snapshot in time, time component
Granularity	atomic – detailed data	summarized data
Summarization level	low, some aggregation possible	high, many aggregation levels
Data model	highly normalized, mostly relational	non-normalized, complex structures, mostly multidimensional DBMS
Transaction type	mostly updates	mostly queries
Transaction volume	high update volumes, low query	periodic loads and summary calculations
Transaction speed	update critical – tuned for updates	retrieval critical
Query activity	low to medium in volume	high query volume
Query complexity	simple to medium	high to very complex
Data volumes	Hundreds of megabytes to gigabytes and up	Hundreds of gigabytes to terabytes and up



Introduction to Data Warehousing

- A **data warehouse** holds data drawn from several data sources, maintained by different operating units within the organization, together with historical and summary transformations.
- The data warehouse is based upon extended database technology to provide the management of the data store. VLDB technology is required.
- The decision making process also requires fairly sophisticated and powerful analysis tools. Two main types of analysis tools have emerged in the last few years: On-Line Analytical Processing (OLAP) tools and data mining tools.
- Data warehousing is an extremely complex subject, an entire course could be devoted to the subject. We will cover enough of the subject to give you some familiarity with the topic and an idea of how they are utilized. In fact, a more recent trend has been toward the *data webhouse* which is a data warehouse which is implemented over a network (the most common being the Internet) with no central data repository.



Introduction to Data Warehousing (cont.)

- Bill Inmon is the acknowledged father of the **data warehouse**. He defines a data warehouse as an integrated, subject-oriented, time-variant, nonvolatile database that provides support for decision making.
 - **Subject-oriented** – the warehouse is organized around the major subjects of the enterprise (such as customers, products, and sales) rather than the major application areas (such as customer invoicing, stock control, and product sales). This is reflected in the need to store decision-support data rather than application-oriented data.
 - **Integrated** – the warehouse houses data from various enterprise-wide sources. The source data is often inconsistent using, for example, different formats. The integrated data source must be made consistent in order to present a unified view of the data to the users.
 - **Time-variant** – the data in the warehouse is only accurate and valid at some point in time or over some time interval. The time-variance of the data warehouse is also shown in the extended time that the data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots.

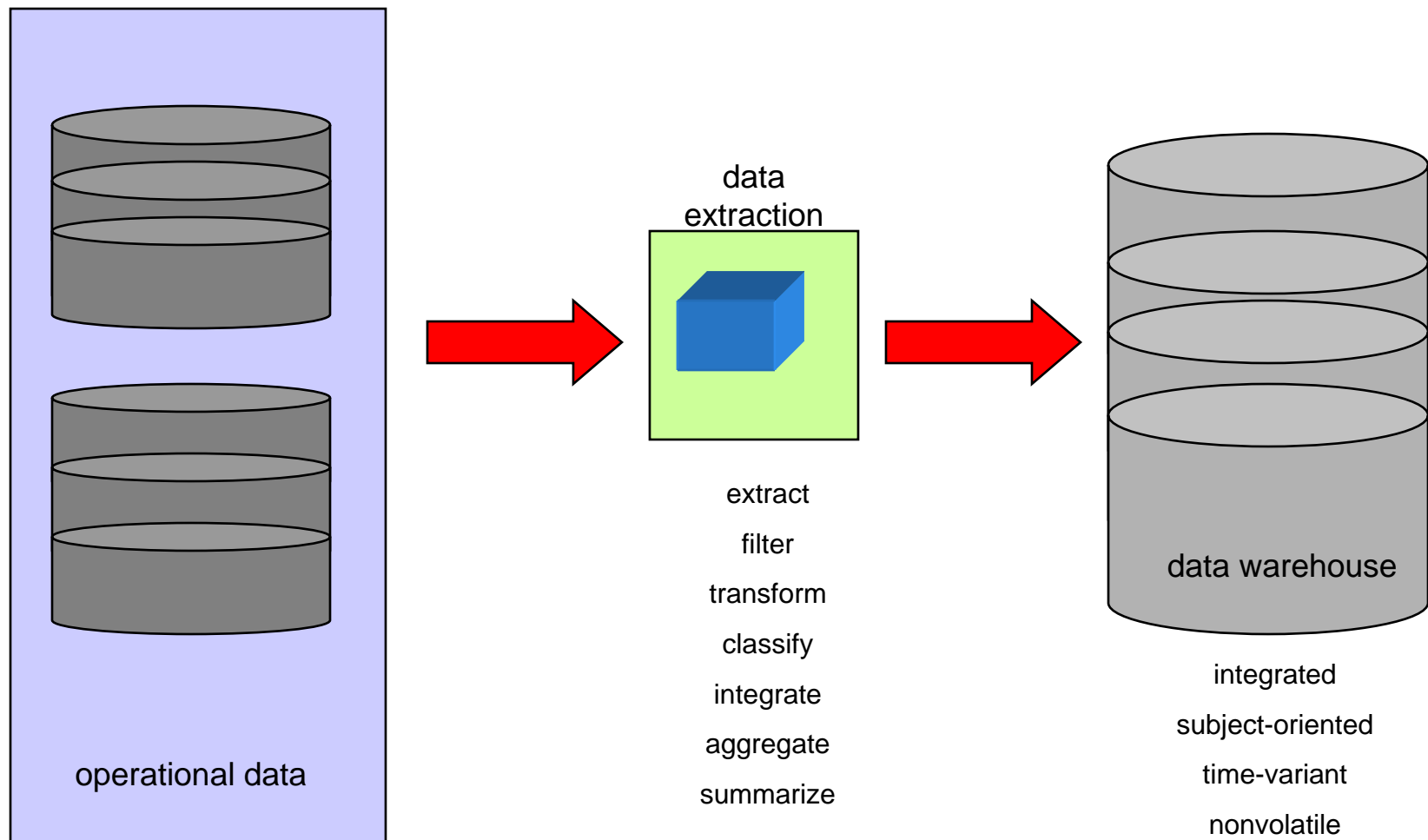


Introduction to Data Warehousing (cont.)

- **Non-volatile** – the data in the warehouse is not updated in real-time but is refreshed from operational systems on a regular basis. New data is always added as a supplement to the database, rather than as a replacement. The database continuously absorbs this new data, incrementally integrating it with the previous data.
- Depending upon who you talk to or which text on the subject you happen to read, you will probably find a slightly different definition of data warehousing. In short, data warehousing is a combination of data management and data analysis technology. Regardless of the definition, the ultimate goal of data warehousing is to integrate enterprise-wide corporate data into a single repository from which users can easily run queries, produce reports, and perform analysis.



Creating a Data Warehouse



Some Issues of Data Warehousing

- While the concept of data warehousing sounds simple enough, there are many problems associated with implementing and maintaining such a system. We'll highlight a few of the more obvious problems in this section of the notes.
- **Underestimation of resources for data loading** – Many developers underestimate the time required to extract, clean, and load the data into the warehouse. This process may account for a significant portion of the total development time, although better data cleansing and management tools should ultimately reduce the time and effort spent on data loading.
- **Hidden problems with source systems** – Hidden problems with the source systems feeding the warehouse may be identified, possibly after years of being undetected. The developer must decide whether to fix the problem in the warehouse and/or fix the source system. For example, when entering the details of a new product, certain fields may allow null values, which may result in entering a null value for such a field even though the data is available and applicable.



Some Issues of Data Warehousing (cont.)

- **Required data is not captured** – Warehouse projects often highlight a requirement for data not being captured by the existing source systems. The organization must decide whether to modify the OLTP system or create a system dedicated to capturing the missing data.
- **Increased end-user demands** – After end-users receive query and reporting tools, request for support from IS staff may increase rather than decrease. This is typically caused by an increasing awareness of the users on the capabilities and value of the warehouse. This problem can be partially alleviated by investing in easier-to-use, more powerful tools, or in providing better training for the users. A further reason for increasing demand on IS staff is that once a warehouse is online, it is often the case that the number of users and queries increase together with requests for answers to more and more complex queries.
- **Data homogenization** – Large-scale warehousing can become an exercise in data homogenization that lessens the value of the data. For example, in producing a consolidated and integrated view of the organization's data, the warehouse designer may be tempted to emphasize similarities rather than differences in the data used by different application areas such as product sales and product inventory.



Some Issues of Data Warehousing (cont.)

- **High demand for resources** – The warehouse can use huge amounts of disk space. Many relational databases used for decision support are designed around star, snowflake, and starflake schemas (these are schemas in which a central schema spawns related schemas which radiate out from the central schema). These schema designs tend to result in the creation of very large fact tables. If there are many dimensions to the factual data, the combination of aggregate tables and indices to the fact tables can require more space than the data itself.
- **Data ownership** – Warehousing may change the attitude of the end-users to the ownership of the data. Sensitive data that was originally viewed and used only by a particular department or business area such as in sales or marketing, may now be made accessible to others in the organization. Indeed, some departments or areas may be unaware of the existence of the warehouse.
- **High maintenance** – Warehouses are high maintenance systems. Any reorganization of the business processes and the source systems may affect the warehouse. To remain a valuable resource, the warehouse must remain consistent with the organization that it supports.



Some Issues of Data Warehousing (cont.)

- **Long-duration projects** – A warehouse represents a single data resource for the organization. However, the building of a warehouse can take up to three years, which is why some organizations are building data marts. Data marts support only the requirements of a particular department or functional area and can therefore be built much more rapidly.
- **Complexity of integration** – The most important area for the management of a data warehouse is the integration capabilities. This means an organization must spend a significant amount of time determining how well the various warehousing tools can be integrated into the overall solution that is needed. This can be a very difficult task, as there are a number of tools for every operation of the warehouse, which must integrate well in order that the warehouse works to the organization's benefit.

