

Section 1: Data warehousing attributes and concepts

Having looked at the historical use of the analysis data and explored some of the factors influencing the evolution of data warehouses, we will now turn to identifying the key attributes of a data warehouse.

It is important to recognize that data warehousing is still an evolving science. As with any evolving technology, particular care must be taken to discount some marketing claims driven by vendors attempting to differentiate themselves from the competitors. For example, the size of the data warehouse should not determine if a data warehouse is really a data warehouse. Some vendor may say that a data warehouse that is only 50 gigabytes is not a full-fledged data warehouse, and they may refer to it instead as a data mart. For a smaller company, 50 gigabytes or even much less can represent every relevant piece of information covering last 10 years and can well represent a powerful data warehouse.

This section explores the data warehousing concepts and attributes. These concepts are grouped into four sub-sections. The first sub-section discusses the reasons for separating the data for business analysis from the operational data. The logical transformation of the data, including data warehouse modeling and de-normalization of the data, are introduced in the second sub-section. Sub-section three reviews the issues associated with physical transformation of the data. Sub-section four discusses the generation of summary views. A very simple and broad definition of a data warehouse follows the discussion of the data warehousing concepts and attributes.

1.1 “Warehousing” data outside the operational systems

The primary concept of data warehousing is that the data stored for business analysis can most effectively be accessed by separating it from the data in the operational systems. Many of the reasons for this separation have evolved over the years. In the past, legacy systems archived data onto tapes as it became inactive and many analysis reports ran from these tapes or mirror data sources to minimize the performance impact on the operational systems.

These reasons to separate the operational data from analysis data have not significantly changed with the evolution of the data warehousing systems, except that now they are considered more formally during the data warehouse building process. Advances in technology and changes in the nature of business have made many of the business analysis processes much more complex and sophisticated. In addition to producing standard reports, today’s data warehousing systems support very sophisticated online analysis including multi-dimensional analysis.

1.1.1 Integrating data from more than one operational system

Data warehousing systems are most successful when data can be combined from more than one operational system. When the data needs to be brought together from more than one source application, it is natural that this integration be done at a place independent of the source applications. Before the evolution of structured data warehouses, analysts in many instances would combine data extracted from more than one operational system into a single spreadsheet or a database. The data warehouse may very effectively combine data from multiple source applications such as sales, marketing, finance, and production. Many large data warehouse architectures allow for the source applications to be integrated into the data warehouse incrementally.

The primary reason for combining data from multiple source applications is the ability to cross-reference data from these applications. Nearly all data in a typical data warehouse is built around the time dimension. Time is the primary filtering criterion for a very large percentage of all activity against the data warehouse. An analyst may generate queries for a given week, month, quarter, or a year. Another popular query in many data warehousing applications is the review of year-on-year activity. For example, one may compare sales for the first quarter of this year with the sales for first quarter of the prior years. The time dimension in the data warehouse also serves as a fundamental cross-referencing attribute. For example, an analyst may attempt to access the impact of a new marketing campaign run during selected months by reviewing the sales during the same periods. The ability to establish and understand the correlation between activities of different organizational groups within a company is often cited as the single biggest advanced feature of the data warehousing systems.

The data warehouse system can serve not only as an effective platform to merge data from multiple current applications; it can also integrate multiple versions of the same application. For example, an organization may have migrated to a new standard business application that replaces an old mainframe-based, custom-developed legacy application. The data warehouse system can serve as a very powerful and much needed platform to combine the data from the old and the new applications. Designed properly, the data warehouse can allow for year-on-year analysis even though the base operational application has changed.

1.1.2 Differences between transaction and analysis processes

The most important reason for separating data for business analysis from the operational data has always been the potential performance degradation on the operational system that can result from the analysis processes. High performance and quick response time is almost

universally critical for operational systems. The loss of efficiency and the costs incurred with slower responses on the predefined transactions are usually easy to calculate and measure. For example, a loss of five seconds of processing time is perhaps negligible in and of itself; but it compounds out to considerably more time and high costs once all the other operations it impacts are brought into the picture. On the other hand, business analysis processes in a data warehouse are difficult to predefine and they rarely need to have rigid response time requirements. Operational systems are designed for acceptable performance for predefined transactions. For an operational system, it is typically possible to identify the mix of business transaction types in a given time frame including the peak loads. It is also relatively easy to specify the maximum acceptable response time given a specific load on the system. The cost of a long response time can then be computed by considering factors such as the cost of operators, telecommunication costs, and the cost of any lost business. For example, an order processing system might specify the number of active order takers and the average number of orders for each operational hour. Even the query and reporting transactions against the operational system are most likely to be predefined with predictable volume.

Even though many of the queries and reports that are run against a data warehouse are predefined, it is nearly impossible to accurately predict the activity against a data warehouse. The process of data exploration in a data warehouse takes a business analyst through previously undefined paths. It is also common to have runaway queries in a data warehouse that are triggered by unexpected results or by users' lack of understanding of the data model. Further, many of the analysis processes tend to be all encompassing whereas the operational processes are well segmented. A user may decide to explore detail data while reviewing the results of a report from the summary tables. After finding some interesting sales activity in a particular month, the user may join the activity for this month with the marketing programs that were run during that particular month to further understand the sales. Of course, there would be instances where a user attempts to run a query that will try to build a temporary table that is a Cartesian product of two tables containing a million rows each! While an activity like this would unacceptably degrade an operational system's performance, it is expected and planned for in a data warehousing system.

1.1.3 Data is mostly non-volatile

Another key attribute of the data in a data warehouse system is that the data is brought to the warehouse after it has become mostly non-volatile.

This means that after the data is in the data warehouse, there are no modifications to be made to this information. For example, the order status does not change, the inventory snapshot does not change, and the marketing promotion details do not change. This attribute of the data warehouse has many very important implications for the kind of data that is brought to the data warehouse and the timing of the data transfer. Let us further review what it means for the data to be non-volatile. In an operational system the data entities go through many attribute changes. For example, an order may go through many statuses before it is completed. Or, a product moving through the assembly line has many processes applied to it. Generally speaking, the data from an operational system is triggered to go to the data warehouse when most of the activity on these business entity data has been completed. This may mean completion of an order or final assembly of an accepted product. Once an order is completed and shipped, it is unlikely to go back to backorder status. Or, once a product is built and accepted, it is unlikely to go back to the first assembly station. Another important example can be the constantly changing data that is transferred to the data warehouse one snapshot at a time. The inventory module in an operational system may change with nearly every transaction; it is impossible to carry all of these changes to the data warehouse. You may determine that a snapshot of inventory carried once every week to the data warehouse is adequate for all analysis. Such snapshot data naturally is non-volatile. It is important to realize that once data is brought to the data warehouse, it should be modified only on rare occasions. It is very difficult, if not impossible, to maintain dynamic data in the data warehouse. Many data warehousing projects have failed miserably when they attempted to synchronize volatile data between the operational and data warehousing systems.

1.1.4 Data saved for longer periods than in transaction systems

Data from most operational systems is archived after the data becomes inactive. For example, an order may become inactive after a set period from the fulfillment of the order; or a bank account may become inactive after it has been closed for a period of time. The primary reason for archiving the inactive data has been the performance of the operational system. Large amounts of inactive data mixed with operational live data can significantly degrade the performance of a transaction that is only processing the active data. Since the data warehouses are designed to be the archives for the operational data, the data here is saved for a very long period.

In fact, a data warehouse project may start without any specific plan to archive the data off the warehouse. The cost of maintaining the data once

it is loaded in the data warehouse is minimal. Most of the significant costs are incurred in data transfer and data scrubbing. Storing data for more than five years is very common for data warehousing systems. There are industry examples where the success of a data warehousing project has encouraged the managers to expand the time horizon of the data stored in the data warehouse. They may start with storing the data for two or three years and then expand to five or more years once the wealth of business knowledge in the data warehouse is discovered. The falling prices of hardware have also encouraged the expansion of successful data warehousing projects.

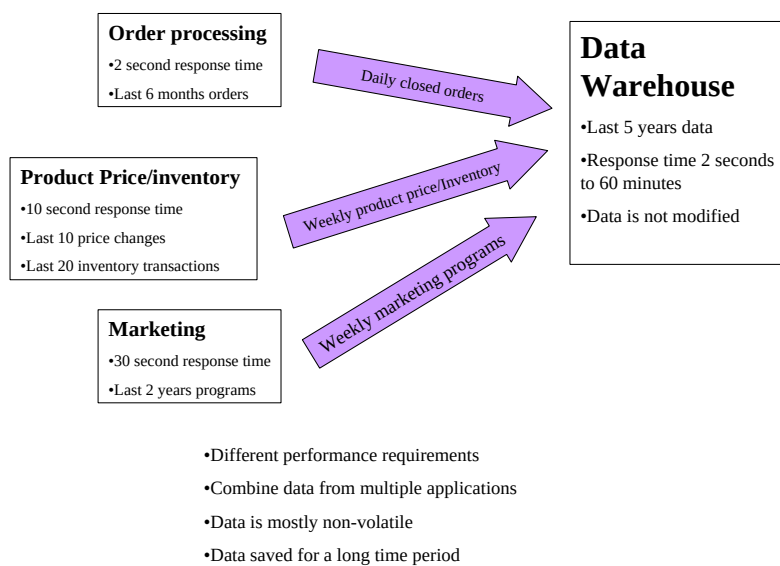
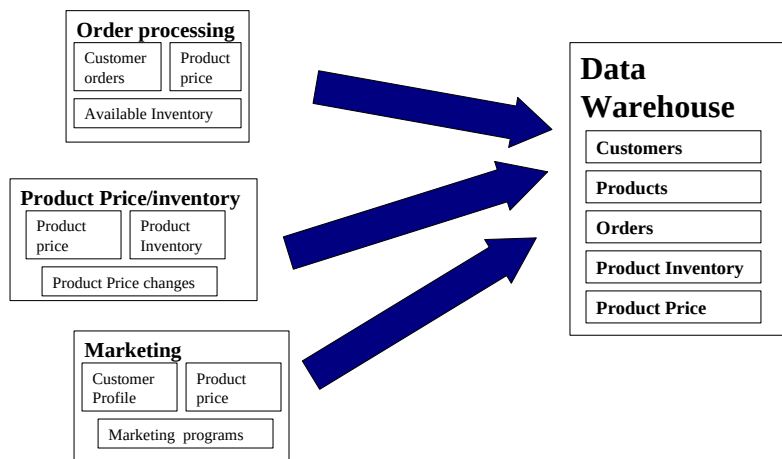


Figure 3. Reasons for moving data outside the operations systems



- No data model restrictions of the source application
- Data warehouse model has business entities

Figure 5. Data warehouse entities align with the business structure