Data Warehousing

by Sophia

WHAT'S COVERED

As organizations have begun to utilize databases as the centerpiece of their operations, the need to fully understand and leverage the data they are collecting has become more and more apparent. However, directly analyzing the data that is needed for day-to-day operations is not a good idea. We do not want to tax the operations of the company more than we need to. Further, organizations also want to analyze data in a historical sense: how does the data we have today compare with the same set of data this time last month, or last year? From these needs arose the concept of the data warehouse. In this tutorial, we will take a closer look at the concept of a data warehouse, and how it informs the decisions made in business environments.

Our discussion will break down as follows:

1. Business Intelligence

A new buzzword that has been capturing the attention of businesses lately is**big data**. Big data refers to such massively large data sets that conventional database tools do not have the processing power to analyze them, as big data sets tend to take up large amounts of storage within the petabyte and exabyte realm. When data reaches these sizes, it becomes much more difficult to analyze and find patterns, as the data continues to evolve. For example, Walmart must process over one million customer transactions every hour.

DID YOU KNOW

One petabyte (abbreviated PB) is equivalent to 1,000,000,000,000,000 (10^15) bytes, or one million gigabytes. One exabyte (abbreviated EB) is equivalent to 1,000,000,000,000,000,000 (10^18) bytes, or one billion gigabytes.

Storing and analyzing that much data is beyond the power of traditional database-management tools. Understanding the best tools and techniques to manage and analyze these large data sets is a problem that governments and businesses alike are trying to solve. **Business intelligence** is used to describe the process that organizations use to take data they are collecting and analyze it. The primary motivation behind businesses seeking to acquire this information lay in the hopes of obtaining a competitive advantage. Besides using data from their internal databases, firms often purchase information from data brokers to get a big-picture understanding of their industries. **Business analytics** is the term used to describe the use of internal company data to improve business processes and practices. **Data mining** is the process of analyzing data to find previously unknown trends, patterns, and associations in order to make decisions. Generally, data mining is accomplished through automated means against extremely large data sets, such as a data warehouse, discussed below.

TERMS TO KNOW

Big Data
Process of handling and analyzing evolving data.

Business Intelligence
Used to describe the process that organizations use to take data they are collecting and analyze it.

business analytics
Used to describe the use of internal company data to improve business processes and practices.

Data Mining
Process of analyzing data to find trends, patterns, and associations in order to make decisions.

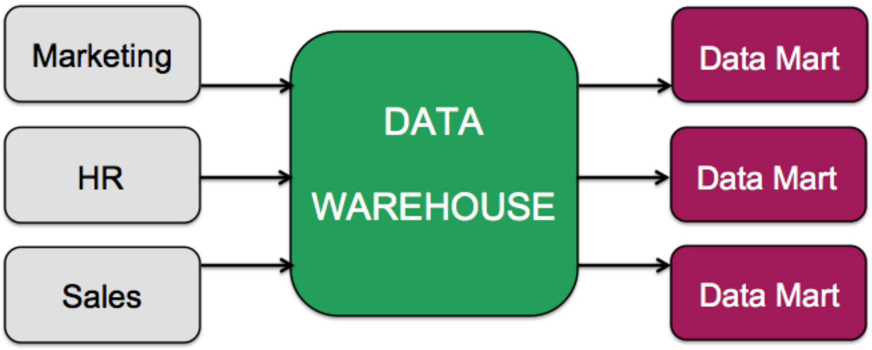
2. The Data Warehouse

A **data warehouse** is a database used by large businesses and organizations for the purposes of collecting and analyzing data related to the business to improve the quality of the business. The concept of the data warehouse is simple: extract data from one or more of the organization’s databases, and load it into the data warehouse for storage and analysis. The goal is to enhance the understanding of the organization's performance in hopes of obtaining a competitive advantage.

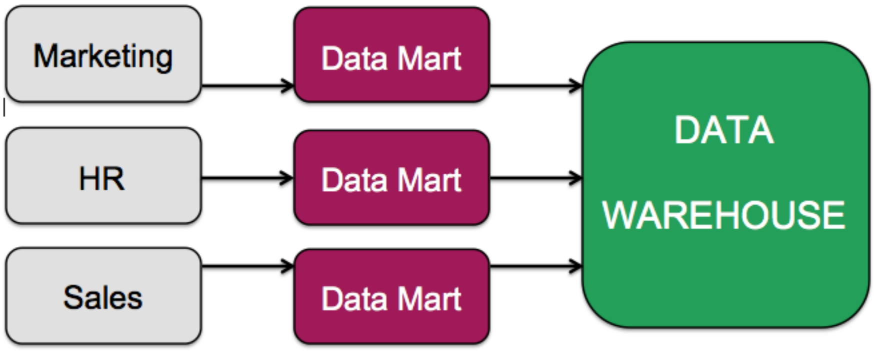
While the concept of the data warehouse is simple, implementation of this concept is not that simple. A data warehouse should be designed so that it meets the following criteria:

- **It uses non-operational data.** This means that the data warehouse is using a copy of data from the active databases that the company uses in its day-to-day operations, so the data warehouse must pull data from the existing databases on a regular, scheduled basis.
- **The data is time-variant.** This means that whenever data is loaded into the data warehouse, it receives a timestamp, which allows for comparisons between different time periods.
- **The data is standardized.** Because the data in a data warehouse usually comes from several different sources, it is possible that the data does not use the same definitions or units. For example, one database might list dates using the mm/dd/yyyy format (e.g., 01/10/2013), while a table in another database might format dates using yy/mm/dd (e.g., 13/01/10). In order for the data warehouse to match up dates, a standard date format would have to be agreed upon and all data loaded into the data warehouse would have to be converted to use this standard format. This process is called extraction-transformation-load (ETL).

There are two primary schools of thought when designing a data warehouse: the top-down approach, and the bottom-up approach. The top-down approach starts by creating enterprise-wide data (data from all departments of an organization, such as marketing, HR, and sales) to form the centralized data warehouse. Then, as specific business needs are identified, this approach creates smaller data warehouses, called **data marts**. The bottom-up approach starts by creating these smaller data marts to solve specific business problems. As these data marts are created, they can be combined into a larger data warehouse.



Top-Down Design



Bottom-Up Design

Each approach has its pros and cons. The top-down approach can be more time-consuming to initially develop, whereas the bottom-down approach takes less time to build and has a shorter initial set-up time. Consequently, the top-down approach has a higher cost at the beginning, but has a lower ongoing development cost, whereas the bottom-up approach has a low initial cost, but doesn't have the benefit of an even lower cost for ongoing development. Aside from time and cost, the benefit from the top-down approach is enterprise-wide data, and the benefit from the bottom-up approach is that the organization starts with specific individual business areas.

TERM TO KNOW

Data Warehouse

A database used by large businesses and organizations for the purposes of collecting and analyzing data related to the business.

Data Mart

A small data warehouse used to solve specific business problems.

3. Benefits of Data Warehousing

To achieve the goal of enhanced business intelligence, and to obtain a competitive advantage, businesses employ the data warehouse. However, businesses and organizations find data warehousing quite beneficial for a number of other reasons. Below is a list of benefits that data warehousing brings to business:

- The process of developing a data warehouse forces an organization to better understand the data that it is currently collecting and, equally important, what data is not being collected.
- A data warehouse provides a centralized view of all data being collected across the enterprise, and provides a means for determining data that is inconsistent.
- Once all data is identified as consistent, an organization can generate one version of truth. This is important when the company wants to report consistent statistics about itself, such as revenue or number of employees.
- By having a data warehouse, snapshots of data can be taken over time. This creates a historical record of data, which allows for an analysis of trends.
- A data warehouse provides tools to combine data, which can provide new information and analysis.

SUMMARY

In this tutorial, you got a chance to see how data plays a major role in the information systems that businesses use to make decisions about the business. A data warehouse is a special form of database that takes data from other databases in an enterprise, and organizes it for analysis. Data mining is the process of looking for patterns and relationships in large data sets. Many businesses use databases, **data warehouses**, and **data-mining** techniques in order to produce **business intelligence** and gain a competitive advantage.

Source: Derived from Chapter 4 of "Information Systems for Business and Beyond" by David T. Bourgeois. Some sections removed for brevity.
https://www.saylor.org/site/textbooks/Information%20Systems%20for%20Business%20and%20Beyond/Textbook.html

TERMS TO KNOW

Big Data
Process of handling and analyzing evolving data.

Business Analytics
Used to describe the use of internal company data to improve business processes and practices.

Business Intelligence
Used to describe the process that organizations use to take data they are collecting and analyze it.

Data Mart
A small data warehouse used to solve specific business problems.

Data Mining
Process of analyzing data to find trends, patterns, and associations in order to make decisions.

Data Warehouse
A database used by large businesses and organizations for the purposes of collecting and analyzing data related to the business.