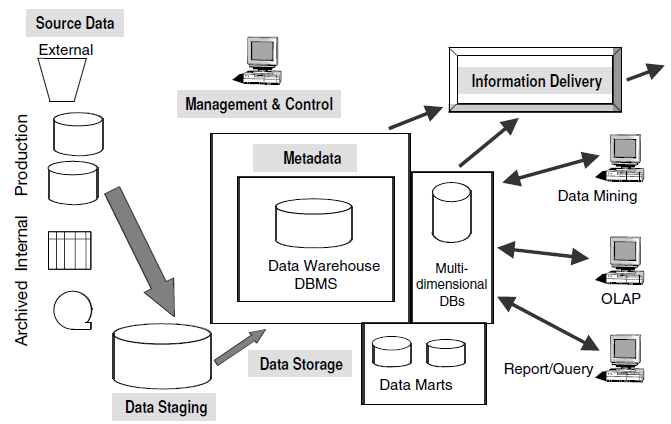
Data Warehouse and Business Intelligence

*Component Overview Architecture of Data warehouse*:



***Production Data***.

1. This category of data comes from the various operational systems ofthe enterprise.
2. Based on the information requirements in the data warehouse, you choosesegments of data from the different operational systems.
3. While dealing with this data, youcome across many variations in the data formats because data resides on different hardware platforms.
4. In operational systems, information queries are narrow. You query an operational sys-tem for information about specific instances of business objects and they are predictable.
5. ***E.g.*** we want Name of single Specified Customer. Or, you may need the orders placed by a singlecustomer in a single week.
6. Our great challenge is to standardize and transform the disparate data from the various productionsystems, convert the data, and integrate the pieces into useful data for storage in the datawarehouse.

***Internal Data***.

1. In every organization, users keep their “private” spread sheets, documents, customer profiles, and sometimes even departmental databases. This is the internaldata, parts of which could be useful in a data warehouse.
2. ***E.g.***Profiles of individual customers become very important for consideration. When your account representatives talk to their assigned customers or when your marketing department wants to make specific offerings to individual customers.

***Archived Data.***

Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files. Sometimes data is left in the operational system databases for as long as five years.

Many different methods of archiving exist. There are staged archival methods.

* At the first stage, ***old***recent data is archived to a separate archival database that may still be online.
* At the second stage, ***older data***is archived to flat files on disk storage.
* At the next stage, the ***oldest data*** is archived to tape cartridges or microfilm and even kept off-site.

***External Data***.

Most executives depend on data from external sources for a high percentage of the information they use. Eg. They use market share data of competitors.

This is not the data from your organization.

For example, the data warehouse of a car rental company contains data on the current production schedules of the leading automobile manufacturers. This external data in the data warehouse helps the car rental company plan for their fleet management.

***Data Staging Component***

After you have extracted data from various operational systems and from externalsources, you have to prepare the data for storing in the data warehouse.

The extracted data coming from several disparate sources needs to be changed, converted, and made ready in a format that is suitable to be stored for querying and analysis.

Three major functions need to be performed for getting the data ready.

* Extract the data,
* Transform the data, and then
* load the data

These three major functions of extraction, transformation, and preparation for loading take place in a staging area. The data staging component consists of a workbench for these functions.

Data staging provides a place and an area with a set of functions to ***clean, change, combine, convert, de-duplicate, and prepare*** source data for storage and use in the data warehouse.

When we implement an operational system, we are likely to pick up data from different sources, move the data into the new operational system database, and run data conversions.

Remember that data in a data warehouse is subject-oriented and cuts across operational applications. A separate staging area, therefore, is a necessity for preparing data for the data warehouse. Now that we have clarified the need for a separate data staging component, let us understand what happens in data staging. We will now briefly discuss the three major functions that take place in the staging area.

***Data Extraction***.

This function has to deal with numerous data sources. You have to employ the appropriate technique for each data source. Source data may be from different source machines in diverse data formats. Part of the source data may be in relational database systems. Some data may be on other legacy network and hierarchical data models. Many data sources may still be in flat files. You may want to include data from spreadsheets and local departmental data sets. Data extraction may become quite complex.

Tools are available on the market for data extraction. You may want to consider using outside tools suitable for certain data sources. For the other data sources, you may want to develop in-house programs to do the data extraction. Purchasing outside tools may entail high initial costs. In-house programs, on the other hand, may mean ongoing costs for development and maintenance.

After you extract the data, where do you keep the data for further preparation? You may perform the extraction function in the legacy platform itself if that approach suits your framework. More frequently, data warehouse implementation teams extract the sourceinto a separate physical environment from which moving the data into the data warehouse would be easier. In the separate environment, you may extract the source data into a group of flat files, or a data-staging relational database, or a combination of both.

***Data Transformation***.

In every system implementation, data conversion is an important function.

Again, as you know, data for a data warehouse comes from many disparate sources. If data extraction for a data warehouse poses great challenges, data transformation presents even greater challenges. Another factor in the data warehouse is that the data feed is not just an initial load. You will have to continue to pick up the ongoing changes from the source systems. Any transformation tasks you set up for the initial load will be adapted for the ongoing revisions as well.

You perform a number of individual tasks as part of data transformation.

* ***Cleaning*** may just be correction of misspellings, or may include resolution of conflicts between state codes and zip codes in the source data,
* ***Standardization*** of data elements forms a large part of data transformation. You standardize the data types and field lengths for same data elements retrieved from the various sources.
* ***Semantic standardization*** is another major task.
  + When two or more terms from different source systems mean the same thing, you resolve the ***synonyms.***
  + When a single term means many different things in different source systems, you resolve the ***homonym.***

E.g.A grocery chain point-of-sale operational system keeps the unit sales and revenue amounts by individual transactions at the check-out counter at each store. But in the data warehouse, it may not be necessary to keep the data at this detailed level. You may want to summarize the totals by product at each store for a given day and keep the summary totals of the sale units and revenue in the data warehouse storage.

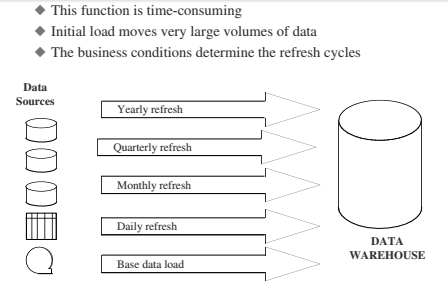
***Data Loading.***

Two distinct groups of tasks form the data loading function.

1. When you complete the design and construction of the data warehouse and go live for the first time, you do the initial loading of the data into the data warehouse storage. The initial load moves large volumes of data using up substantial amounts of time.
2. As the data warehouse starts functioning, you continue to extract the changes to the source data, transform the data revisions, and feed the incremental data revisions on an ongoing basis.

***Data Storage Component***

The data storage for the data warehouse is a separate repository. The data repositories for the operational systems typically contain only the current data. Also, these data repositories contain the data structured in highly normalized formats for fast and efficient processing. In contrast, in the data repository for a data warehouse, you need to keep large volumes of historical data for analysis. Further, you have to keep the data in the data warehouse in structures suitable for analysis, and not for quick retrieval of individual pieces of information. Therefore, the data storage for the data warehouse is kept separate from the data storage for operational systems.

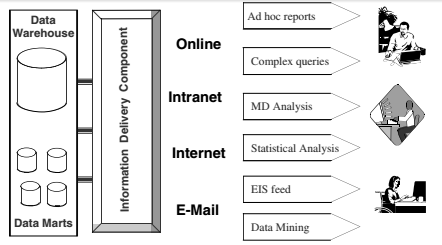


As they are working with the data, the data storage must not be in a state of continual updating. For this reason, the data warehouses are “read-only” data repositories.

***Information Delivery Component***

Who are the users that need information from the data warehouse? The range is fairly comprehensive.

1. The business analyst looks for ability to do complex analysis using the information in the data warehouse.
2. The power user wants to be able to navigate throughout the data warehouse, pick up interesting data, format his own queries, drill through the data layers, and create custom reports and ad hoc queries.



In order to provide information to the wide community of data warehouse users, the information delivery component includes different methods of information delivery.

Ad hoc reports are predefined reports primarily meant for novice and casual users. Provision for complex queries, multidimensional (MD) analysis, and statistical analysis cater to the needs of the business analysts and power users. Information fed into Executive Information Systems (EIS) is meant for senior executives and high-level managers. Some data warehouses also provide data to data-mining applications. Data-mining applications are knowledge discovery systems

***Metadata Component***

Metadata in a data warehouse is similar to the data dictionary or the data catalogue in a database management system. In the data dictionary, you keep the information about the logical data structures, the information about the files and addresses, the information about the indexes, and so on. The data dictionary contains data about the data in the database.

Similarly, the metadata component is the data about the data in the data warehouse.

Metadata in a data warehouse fall into three major categories:

* Operational Metadata
* Extraction and Transformation Metadata
* End-User Metadata

***Operational Metadata***.

As you know, data for the data warehouse comes from several operational systems of the enterprise. These source systems contain different data structures. The data elements selected for the data warehouse have various field lengths and data types. In selecting data from the source systems for the data warehouse, you split records, combine parts of records from different source files, and deal with multiple coding schemes and field lengths. When you deliver information to the end-users, you must be able to tie that back to the original source data sets. Operational metadata contain all of this information about the operational data sources.

***Extraction and Transformation Metadata***. Extraction and transformation metadata contain data about the extraction of data from the source systems, namely, the extraction frequencies, extraction methods, and business rules for the data extraction. Also, this category of metadata contains information about all the data transformations that take place in the data staging area.

***End-User Metadata***. The end-user metadata is the navigational map of the data warehouse. It enables the end-users to find information from the data warehouse. The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

***Management and Control Component***

This component of the data warehouse architecture sits on top of all the other components.

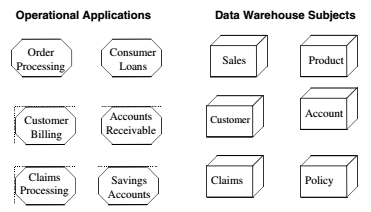
1. The management and control component coordinates the services and activities within the data warehouse.
2. This component controls the data transformation and the datatransfer into the data warehouse storage.
3. On the other hand, it moderates the information delivery to the users.
4. It works with the database management systems and enables data to be properly stored in the repositories.
5. It monitors the movement of data into the staging area and from there into the data warehouse storage itself.

The management and control component interacts with the metadata component to perform the management and control functions. As the metadata component contains information about the data warehouse itself, the metadata is the source of information for the management module.

Bill Inmon, considered to be the father of Data Warehousing provides the following definition: “A Data Warehouse is a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management’s decisions.”

***Subject-Oriented Data***

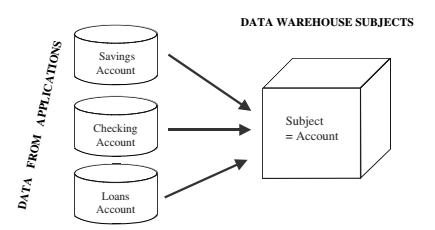
In operational systems, we store data by individual applications. In the data sets for an order processing application, we keep the data for that particular application. These data sets provide the data for all the functions for entering orders, checking stock, verifying customer’s credit, and assigning the order for shipment. But these data sets contain only thedata that is needed for those functions relating to this particular application.



We will have some data sets containing data about individual orders, customers, stock status, and detailed transactions, but all of these are structured around the processing of orders. Similarly, for a banking institution, data sets for a consumer loans application contain data for that particular application. Data sets for other distinct applications of checking accounts and savings accounts relate to those specific applications. Again, in an insurance company, different data sets support individual applications such as automobile insurance, life insurance, and workers’ compensation insurance.

***Integrated Data***

For proper decision making, you need to pull together all the relevant data from the various applications. The data in the data warehouse comes from several operational systems.



Source data are in different databases, files, and data segments. These are disparate applications, so the operational platforms and operating systems could be different. The file layouts, character code representations, and field naming conventions all could be different.

In addition to data from internal operational systems, for many enterprises, data from outside sources is likely to be very important. Companies such as Metro Mail, A. C.Nielsen, and IRI specialize in providing vital data on a regular basis. Your data warehouse may need data from such sources. This is one more variation in the mix of source data for a data warehouse.

***Time-Variant Data***

For an operational system, the stored data contains the current values. In an accounts receivable system, the balance is the current outstanding balance in the customer’s account.

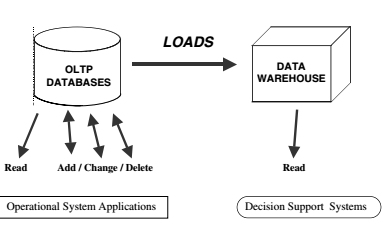
In an order entry system, the status of an order is the current status of the order. In a consumer loans application, the balance amount owed by the customer is the current amount. Of course, we store some past transactions in operational systems, but, essentially, operational systems reflect current information because these systems support day-to-day current operations.

The time-variant nature of the data in a data warehouse

1. Allows for analysis of the past
2. Relates information to the present
3. Enables forecasts for the future

***Nonvolatile Data***

Data extracted from the various operational systems and pertinent data obtained from outside sources are transformed, integrated, and stored in the data warehouse. The data in the data warehouse is not intended to run the day-to-day business.



When you want to process the next order received from a customer, you do not look into the data warehouse to find the current stock status. The operational order entry application is meant for that purpose. In the data warehouse, you keep the extracted stock status data as snapshots over time. You do not update the data warehouse every time you process a single order.