

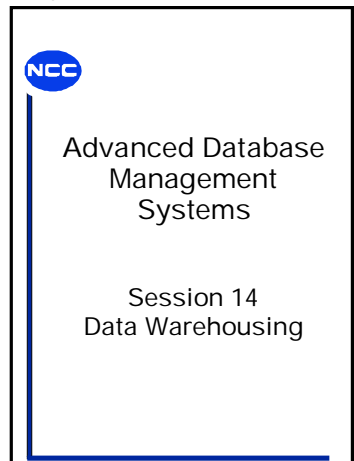
Session 14

Data Warehousing

1 Introduction

V14. 1

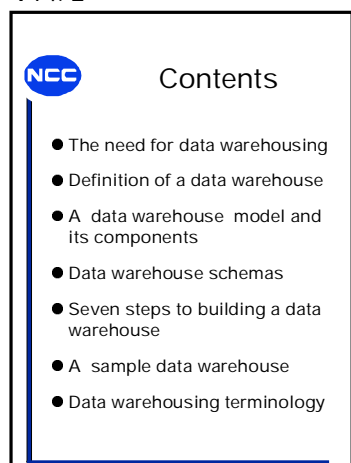
(5 minutes)



Inform students that a handout containing a full set of visuals will be provided to them at the end of this lecture.

1.1 Summary of Topics to be Covered

V14. 2

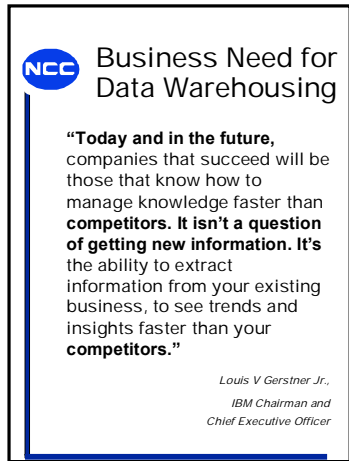


The topics detailed in the visual will be discussed during this session.

2 Need for Data Warehousing

(15 minutes)

V14. 3



NCC Business Need for Data Warehousing

"Today and in the future, companies that succeed will be those that know how to manage knowledge faster than competitors. It isn't a question of getting new information. It's the ability to extract information from your existing business, to see trends and insights faster than your competitors."

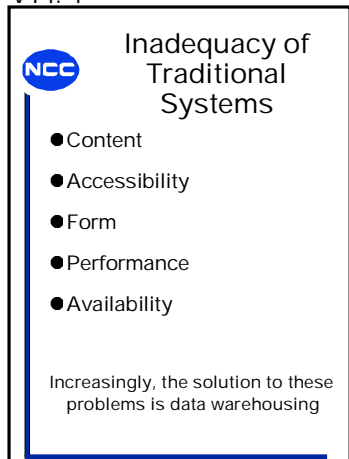
*Louis V Gerstner Jr.,
IBM Chairman and
Chief Executive Officer*

Data warehousing is at the centre of an innovative architecture for information systems in the 1990s. This session will introduce the main concepts and principles of a data warehouse, including its features, architectures and functional components.

The business needs for, and potential benefits of, a data warehouse are typically highlighted by Louis V Gerstner Jr., IBM Chairman and Chief Executive Officer.

"Today and in the future, companies that succeed will be those that know how to manage knowledge faster than competitors. It isn't a question of getting new information. It's the ability to extract information from your existing business, to see trends and insights faster than your competition."

V14. 4



NCC Inadequacy of Traditional Systems

- Content
- Accessibility
- Form
- Performance
- Availability

Increasingly, the solution to these problems is data warehousing

When searching for information to support business decisions in today's dynamic global business environment, many business users find that the traditional sources of data – transaction-based systems – are inadequate. Their inadequacies are manifested in several aspects including content, accessibility, form, performance and availability.

The problems often lie not with the data or its source, but in the limitations of current technology to bring together information from many disparate systems. Increasingly, the solution to these problems is data warehousing.

V14.5

NCC Transformation of “Data” into “Information”

- A data warehouse contains data extracted from one or more systems (data sources)
- A data warehouse provides a multidimensional view of **an organisation’s** operational data to help users make more informed and fast decisions

Data warehouses support information processing by providing a solid platform of integrated, historical data from which various types of analytical information processing are undertaken. Data warehousing provides the facility for integration in a world of *unintegrated* application systems.

A data warehouse is achieved in an evolutionary process. Data warehouses organise and store the data needed for analytical information processing over a long historical time.

A data warehouse contains data extracted from one or more systems (data sources). Examples of these systems include relational DBMSs, object-oriented DBMSs, pre-relational database systems, and legacy systems (file-based data organisations).

A data warehouse provides a multi-dimensional view of an organisation’s operational data to help users make more informed and faster decisions.

3 Definition of a Data Warehouse

(35 minutes)

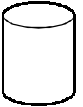
V14.6

NCC Definition of a Data Warehouse

A data warehouse is a

- Subject-oriented
- Integrated
- Time-variant
- Non-volatile

collection of data in support of **management’s decisions**



Data Warehouse

A data warehouse can be defined as “a subject-oriented, integrated, time-variant, non-volatile collection of data” in support of management’s decision making process.

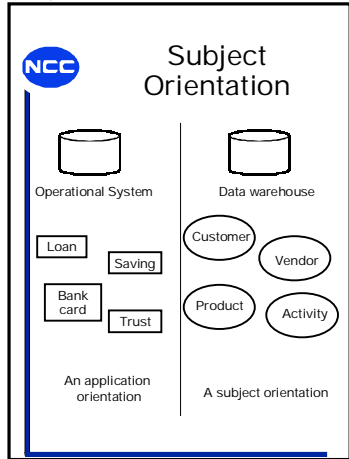
A data warehouse:

- provides a single source of data for all decision support activities;
- is query-intensive in workload;
- is large in size.

The following four visuals will give a fuller explanation of the above definition and highlight some important issues and subtleties underlying the characteristics of a data warehouse.

3.1 Subject Orientation

V14.7



The first feature of the data warehouse is that it is oriented around the major subjects of the organisation. This data-driven, subject oriented approach is in contrast to the more traditional process/functional orientation of applications, which most older operational systems are organised around. An example given in Visual V14.7 illustrates the contrast between the two types of orientations.

The operational system above is designed around applications and functions such as loans, savings, bank cards and trust for a financial institution. The data warehouse, on the other hand, is organised around major subjects such as customers, vendors, products and activities. This alignment around subject areas affects the design and implementation of the data found in the data warehouse.

There are several important differences between these two types of system:

- *System design.* The application oriented operational system is concerned with both database design and process design. The data warehouse focuses on data modelling and database design exclusively. Process design (in its classical form) is not part of the data warehouse environment.
- *Content of data.* Data warehouses exclude data that will not be used for DSS (decision support system) processing, while operational systems contain data to satisfy immediate functional/processing requirements which may or may not be of use to the DSS analyst.
- *Relationships of data.* Operational data maintains an ongoing relationship between two or more tables based on an applicable business rule. Data warehouse data spans over a period of time. Consequently, there are usually many business rules, hence correspondingly many data relationships, represented in the data warehouse between two or more tables.

3.2 Integration

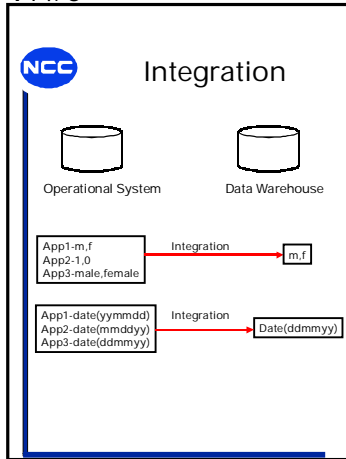
Another important feature of the data warehouse environment is that data contained within the boundaries of the data warehouse is always integrated, with no exception.

This integration is reflected in many ways, including using consistent naming conventions, consistent measurement of variables, consistent encoding structures, consistent physical attributes of data, and so on.

An example in Handout 14.1 highlights the difference between the integration found within a data warehouse and the lack of integration found in an applications oriented operational system.

Issue Handout 14.1.

V14. 8



Over the years different applications designers of the operational system have made numerous individual decisions as to how an application should be built, resulting in differences in many aspects, for example:

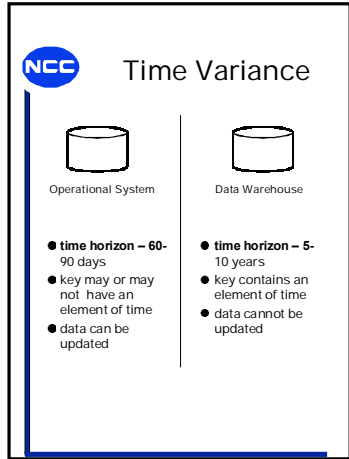
- *Encoding* – Different application designers have chosen to encode the field GENDER in different ways (using M and F, or 1 and 0, or male and female). For the data warehouse, it doesn't matter much how GENDER arrives; "M" and "F" are probably as good as any representation. What matters is that whatever source GENDER comes from, it must arrive in the data warehouse in a consistent integrated state. Therefore when GENDER is loaded into the data warehouse from an application where it has been represented in other than an "M" and "F" format, the data must be converted to the data warehouse format.
- *Measurement of attributes* – Different application designers have chosen to measure pipelines in a variety of ways (in centimetres, or in inches, or in million cubic feet per second, or in yards). Whatever the source, when the pipeline information arrives in the data warehouse it needs to be measured in the same way.

As shown in Handout 14.1, the issues of integration affect almost every aspect of design. Data to be stored in the data warehouse needs to be converted into a singular and *standard* form even when the underlying operational systems store the data differently.

When the DSS analyst examines the data warehouse, the focus of the analyst should be on how to use the data in the warehouse, rather than on wondering about the credibility or consistency of the data.

3.3 Time Variance

V14.9



In an operational system, data is accurate as of the moment of access. In other words, when a unit of data is accessed, it is expected to reflect accurate values at that point in time.

However, the basic characteristic of data in the warehouse environment is very different. Data in the data warehouse is accurate as of some moment in time (i.e. not *right now*). Therefore, data is said to be “time variant” in the data warehouse, as shown in Visual V14.9.

The time variance of data in the data warehouse is demonstrated in several ways:

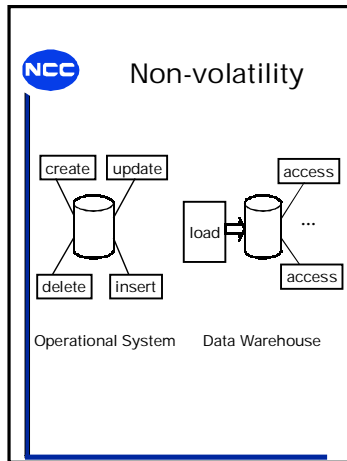
- **Time horizon.** Data warehouse data represents over a long period of time – say from five to ten years. The time horizon represented for the operational environment is much shorter – say from the current values of today up to sixty to ninety days.
- **Key structure.** Every key structure in the data warehouse contains, implicitly or explicitly, an element of time, such as day, week, month, etc. The element of time is almost always at the bottom of the concatenated key found in the data warehouse. Sometimes the time element exists implicitly, for example, when an entire file is duplicated at the end of the month.
- **Update.** Data warehouse data represents a long series of snapshots which, once correctly recorded, cannot be altered or updated (of course the snapshots can be changed if they have been taken incorrectly). On the other hand, operational data, being accurate as of the moment of access, can be updated as the need arises.

3.4 Non-volatility

The fourth characteristic of the data warehouse is that it is non-volatile. Handout 14.2 illustrates this feature of the data warehouse.

Issue Handout 14.2.

V14. 10



In the operational system updates (inserts, deletes, and changes) are done regularly on a record-by-record basis. The basic manipulation of data in the data warehouse is much simpler and there are only two types – the initial loading of data, and the access of data. There is no update of data (in the general sense of update) in the data warehouse as a normal part of processing.

This difference results in some significant consequences. At the design level, the need to be cautious of the update anomaly is no factor in the data warehouse, since update of data is not done. This means that at the physical level of design, liberties can be taken to optimise the access of data, particularly in dealing with the issues of normalisation and physical denormalisation.

Another consequence of the simplicity of data warehouse operation is in the underlying technology used to run the data warehouse environment. Operational processing, which often supports record-by-record updates in an on-line mode, requires complex technology including backup and recovery, transaction and data integrity, detection and remedy of deadlock. Some of this complexity in technology is unnecessary for data warehouse processing.

In summary, the characteristics of a data warehouse – subject orientation of design, integration of data within the data warehouse, time variance, and simplicity of data management – all lead to an environment which is very different from the classic operational environment.

There is another point worth noting. Although the source of nearly all data warehouse data is from operational systems, there is a minimum of data redundancy between the operational environment and the data warehouse environment. This can be explained by the following factors:

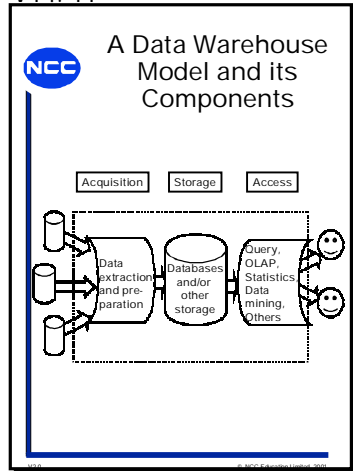
- Data is filtered as it passes from the operational environment to the data warehouse environment. Only that data which is needed for DSS processing enters the data warehouse environment.
- Data in the operational system is very fresh, whilst data in the warehouse is much older. From the perspective of time horizons alone, there is very little overlap between the operational and the data warehouse environments.
- Data warehouses contain summary data which is never found in the operational environment.
- Data undergoes a fundamental transformation as it passes into the data warehouse and most data is significantly altered upon being selected for and moving into the data warehouse, as shown in the diagram in Handout 14.1. It is not the same data that resides in the operational environment from the point of integration.

In the light of these factors, data redundancy between the two environments is a rather rare occurrence.

4 A Data Warehouse (Mart) Model

V14.11

(30 minutes)



There are three basic components in a data warehouse system:

- The Acquisition component.
- The Storage component.
- The Access component.

4.1 The Acquisition Component

1. Identifying required data from legacy systems (and other data sources).
2. Validating that the data is accurate, appropriate, and usable.
3. Extracting the data from the original source.
4. Preparing the data for inclusion into the new environment. This includes the processes of:
 - a) Cleansing – ensuring a satisfactory level of data quality.
 - b) Formatting – establishing the correct format for the data.
 - c) Standardising – ensuring all data is in a consistent form.
 - d) Matching and reconciling – matching up data from two or more disparate sources so that the one *true* value is represented.
 - e) Merging – taking information from two or more sources and consolidating it into one place.
 - f) Purging – eliminating duplicate and erroneous information.
 - g) Establishing referential integrity – making sure that all referential values are accurate.
5. Making the data ready for loading into the warehouse itself.

4.2 The Storage Component

The storage component deals with the creation and management of an environment where the required information is stored, available for user access.

It is important to note that a database (whether it is a relational or not) is only one form of information storage. There have been very large and successful data warehouses which made use of simple flat files, Excel spreadsheets, Web server home directories and object-oriented databases as their storage mediums.

4.3 The Access Component

- *Access using query and reporting tools* – One of the most common forms of warehouse access is using query and reporting tools. At the low end of this category are simple “query manager” products, such as ISQL (Interactive SQL), which allow users to input SQL commands and view the results in tabular form.

At the high end of the category are full-service query management environments such as Business Objects and Cognos. These products provide users with a user-friendly, business-level view of the information they require, while at the same time providing administrators with a suite of tools to manage the environment.

In between are many other query access tools in the marketplace today, including Microsoft Excel and Lotus 1-2-3. These two spreadsheet products, with their backend database attachment capabilities, are often used to support data warehouse access.

- *Access using OLAP tools* – OLAP (On-Line Analytical Processing) tools are a relatively recent innovation and they are basically a specialised form of reporting tool.

Query and reporting tools concentrate on making it easy for users to dynamically request and view information. OLAP tools, on the other hand, give users the ability to interactively and dynamically navigate throughout a large database and explore different aspects of their data warehouse information without having to write queries or format reports.

- *Access for statistical analysis* – Statistical analysis can be considered as the oldest form of data mining. By applying statistical analysis techniques to large volumes of data warehouse data, organisations have been able to optimise performance and achieve operational improvements.

Data warehouses support statistical analysis through end-user spreadsheets such as Lotus 1-2-3 and Excel at the low end, and through powerful statistical analysis packages such as SPSS and SAS at the high end.


- *Access using data mining* – Data mining tools represent one of the newest introductions to data warehouse access. They are used to extract useful

information from the underlying data using various methods of knowledge discovery. This topic is given a separate and detailed treatment in Session 15.

- *Access using graphical and geographic information systems* – The graphical display tools and geographic information systems (GISs) provide some unconventional ways of accessing information and examining problems. With these tools, users can envisage very large and complex multi-dimensional problems in ways which allow them to intuitively interpret the results.

For example, GISs are popular because they allow for the dynamic assignment of numeric data to different sections of a map. These systems present organisations with a bird's-eye view of territories, trouble reports, demographic densities, and many other aspects of a population that would otherwise be very difficult to visualise with normal tabular-form reports.

V14. 12



Data Warehouse Schemas

- Constructing the data warehouse by concentrating separately on simple subsets of organisational data (e.g. departmental data)
- Each simplified schema of departmental data takes the name **"data mart"**
- Multi-dimensional schemas:
 - Star schema
 - Snowflakes schema

There are different types of data warehouse schema. When constructing a data warehouse, concentration should be initially placed on simple sets of organisational data, for example, departmental data. Each simplified schema of departmental data takes the name *data mart*. Each data mart is organised according to a simple structure.

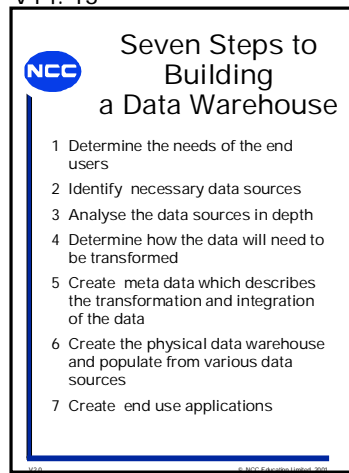
A *multi-dimensional schema* provides a compact and easy-to-understand way of visualising and manipulating data elements which have inter-relationships.

In a *star schema*, a central entity represents the facts on which the analysis is focused, whilst various entities arranged in rays around it represent the dimensions of the analysis.

A *snowflake schema* is an evolution of the simple star schema in which the dimensions are structured hierarchically. It is introduced to take account of the presence of non-normalised dimensions.

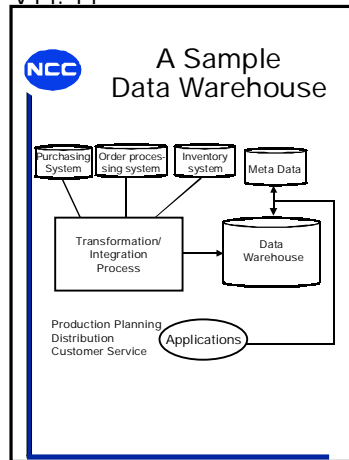
Examples of these data warehouse schemas can be found in books (especially in books 1 to 3) listed in Handout 14.3.

V14. 13



There are some basic steps involved in building a data warehouse as described in Visual V14.13.

V14. 14




The figure shown in Visual V14. 14 outlines the architectural framework for an example data warehouse.

5 Some Useful Terminology of Data Warehousing

(5 minutes)

V14. 15



Some Terminology of Data Warehousing


- **Ad hoc query** – a query whose content cannot be determined in advance
- **Data mart** - small data warehouse focusing on one subject or functional area
- **Data mining** – the process of searching for patterns, trends, or correlations
- **Data staging** – the process of validating, transforming, and integrating data
- **Meta data** – data or information about the entities, such as tables and columns
- **OLAP** – On-Line Analytical Processing
- **OLTP** – On-Line Transaction Processing
- **Transformation** – the act of changing physical data into subject data

- Ad hoc query – a query whose content cannot be determined in advance.
- Data mart – small data warehouse focusing on one subject or function.
- Data mining – the process of searching for patterns, trends, or correlations (see Session 15).
- Data staging – the process of validating, transforming, and integrating data.
- Meta data – data or information about the entities, such as tables and columns.
- OLAP – On-Line Analytical Processing.
- OLTP – On-Line Transaction Processing.
- Transformation – the process of changing physical data into subject data.

6 Summary

(5 minutes)

V14. 16



Summary

- The need for data warehousing
- Definition of a data warehouse
- A data warehouse model and its components
- Data warehouse schemas
- Seven steps to building a data warehouse
- A sample data warehouse
- Data warehousing terminology

Highlight the main points described and discussed in this session.

To gain a better understanding of, and to consolidate their knowledge on the topic, the students are strongly advised to read beyond the core lecture material and consult further readings recommended in Handout 14.3.

Handout 14.1 – Data Is Integrated before Entering the Data Warehouse

