COMPUTER SCIENCE SERIES

COMPUTER DATA MANAGEMENT AND DATA BASE TECHNOLOGY

HARRY KATZAN, Jr.
Chairman, Computer Science Department
Pratt Institute

VNR VAN NOSTRAND REINHOLD COMPANY
NEW YORK CINCINNATI TORONTO LONDON MELBOURNE

COMPUTER DATA MANAGEMENT AND DATA BASE TECHNOLOGY

Library of Congress Catalog Card Number: 74-31147
ISBN: 0-442-24263-8
ISBN: 0-442-23896-7 pbk.

Manufactured in the United States of America

Published by Van Nostrand Reinhold Company Inc.
135 West 50th Street
New York, New York 10020

Van Nostrand Reinhold Company Limited
Molly Millars Lane
Wokingham, Berkshire RG11 2PY, England

Van Nostrand Reinhold
480 Latrobe Street
Melbourne, Victoria 3000, Australia

Macmillan of Canada
Division of Gage Publishing Limited
164 Commander Boulevard
Agincourt, Ontario M1S 3C7, Canada

15 14 13 12 11 10 9 8 7 6 5

Library of Congress Cataloging in Publication Data

Katzan, Harry.
Computer data management and data base technology.

(Computer science series)
Includes bibliographies and index.
1. Data base management. I. Title.
QA76.K342 001.6'42 74-31147
ISBN 0-442-24263-8
ISBN 0-442-23896-7 pbk.

PREFACE

In recent years, there has been an increasing interest in the technical aspects of data base management systems. In fact, many people regularly attend seminars on the subject costing hundreds of dollars. Because of the importance of data base technology and its impact on the computer industry and on enterprises that use computers, basic knowledge of the subject matter should not be limited to those that have the financial means to attend expensive seminars. This is only one reason for this book. Another reason is that the concepts are sophisticated and the methodology complex; they should be recorded in book form for study and reference. Lastly, the topic has academic value since our older concept of data and storage structures has slowly evolved to a consideration of what we actually do with the methodology we have developed.

Thus, the objective of the book is to explore data base technology. In order to fully appreciate the concepts, however, a general background in data management is required—hence the more general title *Computer Data Management and Data Base Technology*. The book is self-contained and only the barest familiarity with computers and data processing is needed. The subject matter is intended and organized for executives, managers, and technical people.

The book is composed of three parts:

I. Fundamentals
II. Data management concepts
III. Data base technology

The first chapter in Part I is an introduction to the concepts of information and knowledge, and how they are used. It sets the tone for the book and is recommended for all readers. Chapters 2 through 4 cover Computer Systems, Computer Software, and Data and Storage Structures, respectively. These chapters reflect the latest concepts in computer technology, and represent the most widely used systems. They can be used to review basic concepts or to refresh one's terminology.

Part II of the book is an introduction to computer data management. Again, the subject matter reflects the latest concepts and includes the following topics: data management concepts and facilities, input and output operations in a data management enviroment, input and output supervision, file organization concepts, and virtual storage access methods.

Part III of the book concerns data base technology and covers: foundations of data base technology, data base structures and representation, descriptive techniques, the DBTG report, the GUIDE/SHARE data base management system requirements, a relational data base model, Integrated Data Store (IDS), and Information Management Systems (IMS).

Lastly, the appendices supplement the text by giving DBTG, IDS, and IMS programs. Appendix A provides a comparison of IDS and DBTG programs; they are reproduced with permission of Mr. Sven Eriksen and the *Honeywell Computer Journal*. Appendix B provides a sample IMS program; it is reproduced with permission of the IBM Corporation.

The book is an outgrowth of a graduate seminar on data base technology given by the author at Pratt Institute. Most of the attendees were practicing professionals, and as a result, the subject matter covered reflects areas of data base technology that are currently of the greatest professional interest.

It is a pleasure to acknowledge the assistance of my wife Margaret who typed the manuscript and provided a liberal amount of assistance.

HARRY KATZAN, JR.

CONTENTS

PART II. DATA MANAGEMENT CONCEPTS

PART III. DATA BASE TECHNOLOGY

PART I FUNDAMENTALS

1 INTRODUCTION

DATA AND REALITY

A well-known computer advertisement reads somewhat as follows: "Not just data. Reality." The implication is powerful. The objective of a computer system is not solely to replace routine clerical tasks—although this is indeed an important function. An equally important function of the computer is to provide insight into the processes of problem solving, decision making, and data analysis. Insight requires information and this fact relates to another important aspect of computer utilization. The computer can be used to store large amounts of data and make it available to a user at a moment's notice. Most computer applications involve a combination of computational and data management operations.

A computer system is deterministic in the sense that the manner in which the system responds to input conditions can be predicted. The set of possible responses is necessarily dependent upon the logical circuitry of the computer and the programs used to control it. Similarly, the data used by the computer during processing operations must be organized and accessed in a well-defined manner. Computer people can be generally grouped into two classes: users and specialists. The user normally utilizes the computer through a language, system, or special hardware device and need not necessarily know the precise details of the structure and operation of the computer system. The engineer, scientist, or analyst that uses the computer with an easy-to-learn language, such as BASIC or FORTRAN, might fall into the "user" category. Two other examples of users are the information specialist that can retrieve information from a data bank with the aid of a query language via a remote terminal, and the retail clerk that uses a point-of-sale device that is connected to a central computer. But because a computer system is deterministic and data must be organized and accessed in a precise manner, someone must know exactly what the computer is doing and this is the role of the specialist. There are specialists in different areas, such as hardware, software, and data management, and there are varying degrees of specialization.

The subject matter of this book is concerned with the specialized topic of data base technology. Before data base technology can be studied, however, a thorough background is needed in computer systems, computer software, and data management. The first two parts of the book are designed to satisfy these basic needs. This review material should provide an introduction to information technology for people who have not had the opportunity of previously being exposed to one or more of the topics. Experienced people can simply omit topics with which they are familiar.

THE NATURE OF INFORMATION

In a society such as ours, one that benefits greatly from the use of mass communications and computer facilities, information is a valuable commodity. In fact, people concerned with the social aspects of computing emphasize that "information is power," and that it can be used as a control mechanism to influence the behavior of individuals. However, major problems arise in information-based societies when information is not up-to-date. The "currency" of information is a problem addressed by data management and data base systems.

In journalism, yesterday's events are without news value. The same philosophy holds true, to some extent, in computer based systems. As depicted in Figure 1.1, the value of information changes with age, where value is generally considered to be a multidimensional attribute. Two major components of the value of information are "quantity" and "quality." The *quantity of information* is measured in terms of volume, completeness, and accessibility. The *volume* of information refers to the capacity of the system and the amount of information that is available for use by a user of the information system. There is a natural limit to the volume of information that a system can store and a user can reference. As shown in Figure 1.1, this limit is reached when the cost of storing and maintaining the information exceeds its value. Some information systems utilize a storage hierarchy concept wherein infrequently used information can be migrated to a relatively inexpensive storage facility.

The capacity of an information system is also related to the efficiency of the system or the accessibility of information, since there exists a relationship between the volume of a storage medium and the speed of access. *Completeness*

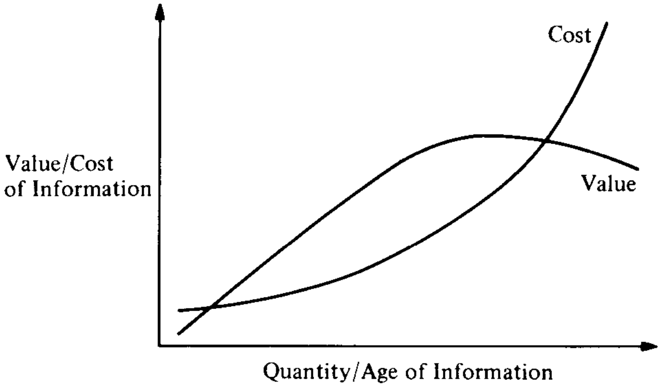

Figure 1.1 The value and cost of information is related to the quantity and age of information, measured in terms of volume, completeness, and accessibility.

refers to the degree to which an information system satisfies the information needs of a user. Completeness can also be transformed into an economic variable since complete information is obviously of greater value than incomplete information but is more costly to maintain. The last attribute of the quantity of information is *accessibility*, which denotes the response time of the system and the facilities available for using it. Systems that provide immediate response to large amounts of information and are also easy to use are naturally more valuable and more costly than systems that provide a lesser degree of accessibility. To sum up, the quantity of information is not an absolute measure, but rather, is a design “trade off” between the value and cost of information.

The *quality of information* implicitly relates to how information can be used and the degree of confidence that can be placed on it. The attributes of quality are timeliness, relevance, accuracy, reliability, and flexibility. It should be noted that these attributes are closely related and both can be used to describe information itself and the information system that makes it available to meet the needs of the user. *Timeliness* refers to the process of collecting, storing, and processing of data and the time factors involved. The key point is that systems that permit informational changes to be made to stored data as related events occur are inherently more timely than systems that require special update procedures.

The *relevance* of information is the measure of how well the information system meets the needs of the user. It can be a measure of the completeness of information or the “degree of preciseness” provided by the query system. *Accuracy* refers to errors in the collecting, storing, and processing of data, and may refer to explicit inaccuracies caused by faulty data or implicit inaccuracies caused by information that is out of date. *Reliability* is an operational characteristic that measures the degree of confidence that the user can place on the information system and the information that it contains. A reliable system provides information that is more timely, relevant, and accurate than an unreliable system. *Flexibility* is the last attribute of information and of an information system, and indicates the diversity of applications for which a given information set can be used. Data that can be used by several users, in several programs, or in making several decisions, is more flexible than data that can be used in a single application. Thus, if data from several applications can be integrated, the quantity and quality of information or of an information system is increased.

The "components of information" given here are not definitive characteristics and should be used to place the subject of information systems in perspective. The various attributes are of greatest value when they are used to describe a particular system or to compare competitive systems.

KNOWLEDGE, INFORMATION, AND DATA

One of the puzzling aspects of our modern computerized society is the fact that even though we are practically submerged in information, we are notably inept at using it effectively.

The world is filled with information. It is inherent in the design of buildings and automobiles, the structure of organizations, and the operations of groups and teams. Yet to a computer or an information scientist, it becomes *data* only when it is recorded on a medium of some kind. It is immaterial whether the media be a notched stick as ancient cave men used to count their wives or sheep, or a modern device such as a punched card or magnetic tape. Informally, *information* may be regarded as raw or processed data used for making decisions, although an information scientist would regard it as that which is generated by a change in a unit of storage. Either definition is satisfactory since the notion is rather well-defined anyway. *Knowledge* implies organization and is defined as the systematic organization of information and concepts. Knowledge can also be defined as the assignment of meaning to information–and therein lies the difficulty for organizations using computer-based information systems.

What is meaning? Clearly, it is a process of naming and identification. But it is more than that. It is also the identification of an object or event by name as a result of a common, agreed upon correspondence between an *event* metaphor and a *name* metaphor. It follows that the meaning that we assign to an event metaphor is implicit in our response to it.

One might inquire at this point what all of these definitions have to do with information systems. The answer is, essentially, that intelligence is a behavioral property and that an information system which we can use intelligently is one which takes account of the context in which it operates. One might also state that this type of information system must be adaptive and this is certainly true. But the basis of an effective information system is deeper than that. *Intelligent behavior,* on the part of a man, a machine, or both, *is the detection of the change* *of meaning brought about by a shift of context.* Therein lies the foundation of a useful man-machine information system.

This latter point must be amplified. Many information systems store data and a subset of this data is eventually used for making decisions or preparing reports. For the most part, however, data is single-purpose in that it is collected and used with a small number of objectives in mind. In a large organization, like the City of New York, a significant amount of data is required to support its diverse activities and much of it is redundant. Using the notion of intelligent behavior given previously, a minimal amount of data would be stored with its specific meaning being dependent upon the context in which it is used. An accepted name for this would be a *common data base.*

SYSTEMS CONCEPTS

The preceding discussion of knowledge, information, and data emphasizes the fact that before a data item can be made usable, a relationship must be established between it and an independent entity, such as a user, an operational environment, or another data item. This is the process that ascribes meaning to data.

Essentially, we are talking about a system. In his book *A Methodology for for Systems Engineering,* Arthur D. Hall defines a system as follows:

> A *system* is a set of objects with relationships between the objects and between their attributes.

Objects are the components or parts of the system and in a general sense are unlimited in scope and variety. In actual practice, however, the objects of which a system is synthesized give the system its structure, and implicitly determine the practical limits on its functional operability. Attributes are properties of objects, such as the resistance of a wire or the age of an individual, and normally correspond to data stored in an information system. Relationships connect the system together, so they can be regarded as a single entity, and can take the form of a physical connection, a logical similarity, a casual rule, and so forth. In information systems, relationships are used to connect data items to form data aggregates.

Systems exist with the support of an environment, and in most cases, the environment determines the nature of the system itself. The *environment* of a system exists as the objects outside of a system, whose attributes affect the system or are changed by the operation of the system. Open systems interface or exchange information with their environment; closed systems do not interface with their environments and there is no exchange of energy between them. In the same manner that objects possess attributes, systems do also. Systems have been variously classed as being adaptive, probablistic, deterministic, stable, and possessing a feedback mechanism.

The nature of the existence of systems is referred to as *systems ontology,*