THE UNIVERSITY OF MANITOBA

AN ARCHITECTURE FOR A DATA MANAGEMENT SYSTEM

by

JOHN W. REECE

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

WINNIPEG, MANITOBA

OCTOBER 1977

AN ARCHITECTURE FOR A DATA MANAGEMENT SYSTEM

BY

JOHN W. REECE

A dissertation submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

ACKNOWLEDGEMENTS

The author gratefully acknowledges the aid and encouragement received from his advisor, Professor R.J. Collens. He also wishes to thank S.A. Bukhari and H. Ferch for their suggestions to improve the clarity of the text. Their contribution, together with that of Mrs. L. Burkowski, was particularly important because of the short time-frame. For the preparation of the manuscript, he thanks Richard McDonald.

The research was financially supported by the Canadian Wheat Board. The author thanks M. Head and F. Jefferson for making the funding possible.

Finally, the author cannot thank his wife enough for her consistent support and baby-sitting throughout preparation of his thesis.

TABLE OF CONTENTS

1 INTRODUCTION

This thesis describes the architecture of a data management system DATA RESOURCE MANAGER (or DRM). The term data management system(DMS) is chosen to designate a complete system of storage, manipulation, definition, and control facilities rather than the simpler physical data base management system (DBMS) which provides the accesing methods of the DMS. DRM was designed to make a high-level data language, useable by non-programmers, available with current data base organizations. In addition, it illustrates that the methodologies chosen to implement these facilities lead to a flexibility which allows migration from current to more advanced technologies as the opportunity arises.

Chapter 2 introduces the data base concept and identifies some of the problems that have limited the success of data base systems. Historical development of the concept is explored from both organizational and technological viewpoints. Three significant emerging trends discussed are:

- management of data as a resource

- defining data in terms of its information content
- a set-theoretic (relational) model of data structure and manipulation

The fundamental purpose of DRM - the goal of developing a technology to complement and exploit these trends - is defined.

Chapter 3 defines objectives for data management systems within a context of management objectives for data. The characteristics of DRM are described and related to the objectives. Where possible, significant alternatives are compared to clarify the selection rationale.

Chapter 4 describes the DRM architecture in detail. Success of a data management system depends upon a powerful end-user language. The Data Resource Access Facility (DRAFT) is introduced. Examples of all types of statements in the stand-alone, interactive part of DRAFT are given. Overall module structure and control flow gives an overview of the system internals. The specific implementation described is a stand-alone

query language over an IMS data base, although the system is equally adaptable to statements imbedded in a batch host language or a differemt data base system. Function and operation of the major modules of DRM are specified. Numerous examples are used to clarify the request translation and execution process.

Chapter 5 summarizes the insights gained from the development of the DRM system architecture. Areas for further study and development of the DRM concept are identified.

2 DATA MANAGEMENT SYSTEMS

There has been a steadily widening interest in data base systems over the last decade. Many corporations have made large investments in data base technology but, alas, few have been rewarded with even a fraction of the benefits attributed to the concept. This chapter examines some of the reasons for this failure and the potential rewards to those organizations remaining optimistic of final success.

2.1 BACKGROUND

The main problem in discussing the data base concept is the lack of common agreement as to what a data base is. (There isn't even agreement to the spelling of the word - see Bibliography.) As a result there has been a divergence of expectations from the reality of technology. Advances in both management's understanding of data and the technology of data base systems will eventually lead to the expected benefits.

To set the scene for the following discussion several terms are now defined. For this thesis a DATA BASE is defined thus:

> A DATA BASE is a collection of interrelated data, stored with controlled redundancy, independent of application programs, to serve multiple requestors. The data base is managed by a single software system (known as a data management system).

Three levels of structures have been defined for

data bases. They are commonly known as <2>:

Internal - the physical structure as stored on disk or mass storage

Conceptual - the overall logical structure as defined by the DBA. (The data base as perceived by the user community.)

External - the subset of the data base defined by a particular user or program.

Three roles are often referred to:

Requestor - Any terminal user or application program which accesses the data base.

Data Administrator - Has overall responsibility for management of integrated data. The DA defines policy and formulates the long-range data base plan.

Data Base Administrator - Is responsible for operational management and control of the data bases. The DBA defines data bases, protects data integrity, and optimizes data base performance.

A widely held misconception has data base systems

maintaining a centralized pool of all enterprise data forming the basis of a total Management Information System. It is likely beyond human capability even to fully understand the scope of the undertaking <15>. There is doubt that it would be desirable to construct a total MIS, were it possible. Decision making is too abstract to identify the inputs to a given decision with certainty. At best the data base can be a source of some of the information to support decision making.

Simply giving the manager more information is not going to improve his decisions. The need is to provide an effective information delivery system. Conventional application files contain much of the data the manager needs. Often this data is unavailable to management due to a lack of relatability of one file to another. The time and cost of developing programs to relate the data may be too great to bear. On the other hand, the information the manager receives is buried in too much data.

There is an emerging trend to treat data as a valuable resource which can be managed by well known principles to maximize its potential value to the

enterprise. Proponents of this concept, known as Data (or Information <40>) Resource Management <37> suggest that achievable expectations for data are that it be:

- manageable like any other resource
- organized to facilitate ad-hoc requests
- capable of systematic growth and adaptation
- integrated across organizational units of the enterprise

With these expectations in mind, chapter 3 develops DRM objectives and requirements in top-down fashion.

In parallel with the evolution of the understanding of data, data base technology has developed from its predecessors: generalized input/output routines and file access methods. Mostly DBMSs have been used as little more than sophisticated access methods and data bases as no more than complex files. Each system is designed in the traditional manner. Data bases are structured in the limited context of the application. Unrelated data bases, with application orientation, have no advantage over conventional files.

To integrate data across several applications, a new approach to analyzing information requirements is essential. The entity concept structures data in terms of the information it represents. This idea is explored in the next sub-section.

The first commercial data base management systems appeared around 1968. Systems in wide use today date from between then and 1972. They are characterized by the data structure that the user sees. They implement one of two models: network<13,8> or hierarchic<28,34>.

In the early seventies, a third structure called the relational model was developed. It possessed a greater conceptual simplicity than the others and had a firm foundation in mathematical set theory. There has been much academic interest in relational data bases since Codd first solidified the concepts in 1971 <12>. A significant body of knowledge exists about the properties of relations and operations upon them. The few implementations <27,30,35> are experimental and use sophisticated physical data structures and access methods. This single technical limitation is a major

factor slowing the acceptance of relational systems.

To date no commercial relational data base system is available. Recently emphasis has shifted from the rigorous definition of properties to practical aspects of implementing a relational data management system. This trend has lead to the DRM concept. Advantages of a relational approach are identified in subsection 2.1.2.

2.1.1 THE ENTITY CONCEPT

Traditionally data structure has been drawn from the design of the application system itself. To develop an integrated, shareable data base, a new source of data organization must be found. The data base is not merely a repository of values, but is a model of some limited universe. Thus our understanding of that universe might be used as the basis for determining data requirements.

The entity concept is a way to look at the nature of information itself. There are three realms when

looking at information and, as Engles <19> noted, we tend to jump from one realm to another without warning. A clear understanding of these realms and their relationship to one another is needed in order to systematize their mapping in the data base.

Firstly there is the real world of objects, people, concepts, etc. which are of interest. These "things" are known as Entities. Employees, products, and bank accounts are entities. Properties are the characteristics of an entity - age, colour, or balance.

The information realm consists of ideas about the real world which exist in the minds of men and women. Entities are represented by property classes called attributes. That Jane Bloggins is a programmer would be represented by a value "programmer" in her POSITION attribute. Our conception of an entity is represented by a collection of attributes, which correspond to the properties of the entity, grouped together to form an Entity Record. A property by which an entity is commonly known (ie "Jane Bloggins") is called an ID attribute. Relationships between entities are also of

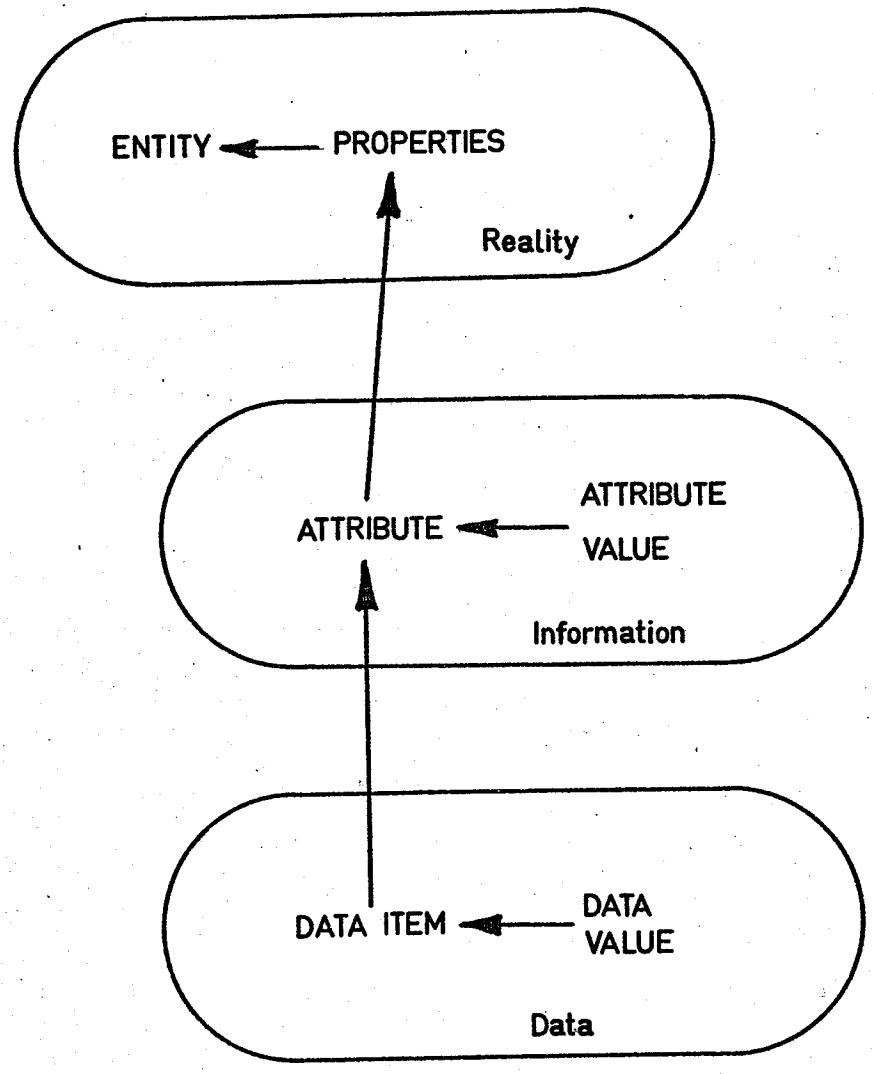

FIG. 2.1 THREE REALMS OF INFORMATION
(after Engles)

interest. The records and relationships relevant to a requestor form an external (or logical) view of the data base.

In the third or data realm are the data items that represent attributes and data values which represent attribute values. There may be many equivalent codings for a given attribute value (ie "2", "TWO", "II"). Values of the ID attribute are represented by key items which are used to access instances of entity records.

Mapping from reality to information is a matter of determining the things of concern to the enterprise and defining entities with attributes corresponding to the properties of interest. The set of these entities which are computerized forms the conceptual data base. This process is largely a matter of judgement on the part of the data base planners. Some care must be exercised in analyzing what is really of interest to an enterprise and, therefore, what the entities are. For instance, is there one PERSON entity in a company's data base or separate EMPLOYEE, SPOUSE, and CHILD entities? (What are the implications of this decision if a father and

son both work for the company?). A set of rules to aid in making these decisions, known as Normalizations, are discussed in the next sub-section.

The conceptual data base is defined to the DMS by statements in the Data Definition Language (see section 4.1). The data realm consists of the corresponding data items and records on physical storage. Mapping between the two realms is defined by the String Model (section 3.3.4). Conceptual records may be materialized or used to form a basis for definition of external views of the data base. Storage and maintenance of physical data and materialization of external records from data is the responsibility of the Data Management System.

2.1.2 THE RELATIONAL MODEL

In its simplest form the relational model is a tabular representation of data. A table corresponds to an entity. Each row is an instance of the entity and each column is an attribute of the entity record. For example the EMPLOYEE relation shown as a table in

EMPLOYEE

NAME	POSITION	DEPT	SAL
ROSS	PROGRAMMER	4	125
DOE	LIBRARIAN	4	100
LEE	PROGRAMMER	4	150
SMITH	CLERK	9	80

DEPARTMENT

DEPT	MANAGER
9	WILLIAMS
2	JONES
4	HANSEN

E1 List all employees who are programmers.

NAME	POSITION	DEPT	SAL
ROSS	PROGRAMMER	4	125
LEE	PROGRAMMER	4	150

E2 List all employee's names and salaries.

NAME	SAL
ROSS	125
DOE	100
LEE	150
SMITH	80

E3 List all employees and their department.

NAME	POSITION	SAL	DEPT	MANAGER
ROSS	PROGRAMMER	125	4	HANSEN
DOE	LIBRARIAN	100	4	HANSEN
LEE	PROGRAMMER	150	4	HANSEN
SMITH	CLERK	80	9	WILLIAMS

FIG. 2.2 EXAMPLES OF OPERATIONS ON RELATIONS

Figure 2.2, where each row is a person and the columns specify the employee's name position, etc.

The approach has a set of simple but very powerful operators to query and manipulate tables and entries. Results of all operations are, themselves, tables which may be displayed or further operated upon. In practise the most common operators are: Selection, Projection, and Joining.

Selection extracts a subset of the rows of a table which satisfy a given condition. Example E1 shows the result of selecting all programmers from the EMPLOYEE relation.

Projection extracts a subset of the columns from a table as specified by a list of column names. Example E2 shows the projection of the NAME and SAL columns from the EMPLOYEE relation.

Join concatenates rows of two tables which share a common value in a column of both tables. In example E3 rows of EMPLOYEE and DEPARTMENT are joined when their