DEVELOPMENT OF AN EXPERT SYSTEM FOR DBMS SELECTION

by

David A. Van Rossum
Boeing Computer Services Company

INTRODUCTION

Many of the projects being developed by Boeing Computer Services Company in support of The Boeing Company and commercial contracts, call for the development or design of a data base system. Although Boeing does develop data base management software, it also relies heavily on software offered commercially. How well a data base management system (DBMS) satisfies the needs of a data base development project has a significant impact on the performance and operational costs of the final system. Therefore, great care must be exercised to ensure that the DBMS chosen is the best one for the information system being proposed.

Selecting the right DBMS requires the presence of a DBMS expert, one who is knowledgeable about the characteristics and strengths of the DBMS' currently being offered. This is a critical problem with most companies and organizations. A person or group of personnel having broad expertise on many DBMS' is very difficult to maintain. Also, once DBMS selection has occurred the need for general DBMS expertise is reduced drastically while the need for expertise on a specific DBMS increases.

DBMSCon provides the general DBMS expert. It will, with a brief but comprehensive session, develop a list of DBMS' that should be considered. Depending on the data base requirements, the list of recommended DBMS' may be large, small, or empty. If it is empty, there is no DBMS within DBMSCon's knowledge base that would adequately satisfy the requirements. If there are several in the list, they are presented in descending order of preference with certainty factors assigned to indicate the likelihood of a particular DBMS' capability to fulfill the requirements. It uses the "knowledge and expertise" of the DBMS expert to provide solutions during the selection of a DBMS. Its reasoning structure is available to the user during a session. In the following paragraphs, we describe the characteristics of DBMSCon.

Although DBMSCon is not a fully operational expert system at this time, the intent is to continue to improve and expand its capability to the level of an objective DBMS consultant.

PREPARATION

DBMSCon is not a shortcut for doing requirements or logical design of an information system. DBMS selection should not occur until these two steps have been completed. Extensive research will be necessary to supply DBMSCon with accurate and conscientious responses. Information collected during requirements definition and logical design will be utilized by DBMSCon to recommend not merely a single DBMS but several DBMS' that most closely satisfy the system requirements. The degree of certainty of the DBMS selected is also dependent on the users' certainty of the information supplied (confidence in the completeness of the requirements and logical design for the system).

There is no reason DBMSCon could not be used during the requirements and logical design phases to perform "what if" scenarios and obtain preliminary analyses. DBMSCon identifies and provides valuable intermediate information that may be used during the aforementioned phases.

USER INTERFACE

The user has a dialogue with DBMSCon. During the dialogue, DBMSCon asks the user various questions about the proposed information system and its environment. With each response from the user, DBMSCon builds its knowledge base and draws interim conclusions. When it comes to the point where more knowledge is required, it asks the user for it. If the user cannot supply the knowledge, DBMSCon will attempt to ascertain the information indirectly via secondary lines of questioning. If this fails, DBMSCon will continue refining its DBMS selection without the desired data. This does, however, reduce the certainty in selecting the best potential DBMS' for the application.

DBMSCon checks for valid entries. If a nonvalid entry or a "111 is given in response to a question, DBMSCon will display the list of valid responses it is expecting. The user need only type in as many characters as is necessary to distinguish the response from the other valid responses. If a response is misspelled, it is automatically corrected, if possible, and accepted.

DBMSCon is always ready to explain its reasoning process to the user. If the user feels that a question is not germane to the problem or that DBMSCon is not pursuing the right line of knowledge acquisition, he/she can respond "WHY" to any question. DBMSCon answers by first repeating the users question, for example:

"WHY IS IT IMPORTANT TO DETERMINE THE SIZE OF THE DATA BASE?"

It then informs the user of what it currently knows, followed by what is necessary to make a particular conclusion. If the user still feels that the conclusion it is trying to establish is
not germane, he/she can ask "WHY" again. DBMSCON will explain the new conclusion it can make after establishing the one just explained. Further "WHY'S" will cause DBMSCON to display its chain of reasoning, thus assuring the user of the reasonableness of its thought process.

The user may also probe forward in DBMSCON's chain of reasoning by asking "HOW" it is going to establish a particular conclusion. Asking a series of "HOW" questions causes DBMSCON to chain forward to display its reasoning process.

The "HOW" and "WHY" questioning capability gives the user the same probing capabilities with DBMSCON that he would have with its human counterpart. It helps to establish a factor of confidence for the DBMSCON user. The language used to display this reasoning process is easy to understand. It is a translation beyond the more tersely formatted language used by the knowledge engineer in building DBMSCON's knowledge base.

Finally, DBMSCON's conclusion is a list of DBMS' in descending order of applicability to the proposed data base system. Following each DBMS is a "certainty factor" between -1.000 and 1.000 which is an indication of DBMSCON's confidence that the DBMS can fulfill the data base requirements.

Following the conclusion, the user is given the ability to question DBMSCON regarding any of the DBMS' (not) recommended. The user can trace any set of reasons why a particular DBMS appeared or did not appear. In its research state, DBMSCON allows the user to change any of the knowledge elements that led to erroneous conclusion, to further strengthen certain conclusions, or to add elements of information that DBMSCON did not have. In its protected state, changes will be made only by the knowledge engineer, or by the DBMS expert.

The capability to easily change or add to DBMSCON's knowledge base and to question its reasoning process gives DBMSCON the dynamism of its human counterpart. If the user is not a DBMS expert and is seeking advice from DBMSCON, DBMSCON will provide expert advice. If the user is a DBMS expert, DBMSCON can act as an expert colleague to confirm or enhance the users independent conclusions. It is an excellent source of second opinions and is being tested with past and current case studies.

**KNOWLEDGE ENGINEER INTERFACE**

Expert systems are built by knowledge engineers. They are tested and enhanced by experts, but built by knowledge engineers adept at extracting the knowledge from experts(s). The extracted knowledge is formatted into "rules" and stored in the knowledge base. The knowledge engineer, in conjunction with the expert(s), must define the allowable set of responses, determine what should be done if a user does not have an answer or supplies several, and determine certainty factors for the results of each rule.

It is the DBMS expert's responsibility to formulate tangible axioms and determine their relationships and value in the scheme of an expert system. The intricately acquired understanding of relationships an expert may have must be extracted and presented, compound ideas uncoupled, and discrepancies resolved. The logic is accumulated through intensive interaction between one or more experts in the data base field and the knowledge engineer. The logic built by the knowledge engineer can only be as good as the expert(s) used to develop the expert system.

S.1, a software package developed by Teknowledge, Inc., is the expert system development tool used by DBMSCON to provide the knowledge engineer the capability to load, test and debug the rules to be installed in the knowledge base (the expert system). In addition, S.1 provides the ability to load, modify and re-execute a set of test cases that can be retained as baselines for evaluating modifications to the expert system. S.1 executes on a VAX in a VMS environment.

S.1 can also execute on a Xerox 1108 LISP machine using the facilities of Interlisp-D, a software package that provides extensive windowing for debugging purposes. The knowledge engineer uses the capabilities of S.1 to develop rules in the following typical format:

**Premise:** data characterize = hierarchical

**Conclusion:** db. structure = hierarchical <60>, network <30>

The words used in the above rule are defined by the knowledge engineer. Complete phrases can be defined with a single definition. Terminology does not need to be completely defined before development of an expert system begins. This permits the knowledge engineer the flexibility of easily expanding the scope of an expert system.

The certainty factor (the values in brackets in the above rule) is a variable determined by the expert/knowledge engineer to define the relative certainty of a conclusion based on a given premise. A certainty factor is established for each value in a conclusion, and is evaluated based on the single premise as if it were standing alone. Premises may use most any level of boolean sophistication, including parenthetical nesting and NOT, in its qualification for a conclusion.

Certainty factors are not combined in a statistical manner. The evaluation of information by an expert is not totally based on statistical evaluations or historical probabilities, but on a true knowledge of the environment and needs. As a result, a method of combining certainty factors other than through basic statistical means was developed. The
method used in S.I. is identical to those developed for some of the early expert systems.

THE REALM OF DBMSCON

The set of rules currently defined in DBMSCON encompasses a complex and diverse range of subject areas. Many of the responses to the questions presented to the user take more than casual investigation. This does not necessarily limit the user during an analysis session as he/she has the option to by-pass questions or experiment with them in order to obtain a cursory set of "preferred" DBMS'. But, conscientious responses must be made for the final analysis in order to provide a valuable decision process.

There are three primary areas of influence used by DBMSCON to select the proper DBMS: DBMS CHARACTERISTICS, DATA BASE STRUCTURE, and HARDWARE. The primary areas are addressed in the first level of a hierarchy of logic used by DBMSCON (shown below). As indicated by a hierarchy each of the primary areas are determined independent of one another. The primary areas can also be individually reviewed by a user independent of the final DBMS recommendations. DBMSCON analyzes conclusions in each of the three primary areas and combines them (via rules and certainty factors) to recommend the DBMS'.

![Diagram of DBMS Selection Hierarchy]

Each of the primary areas can be divided into secondary, subject areas in a hierarchical fashion to whatever depth is necessary to satisfy the needs of the expert system (see Figure 1). Rules may be defined at any level or node in the hierarchy (i.e., vertically within the hierarchy). Subjects may cut across the hierarchy (horizontally). An example of this would be data volume. There are limitations in some DBMS' concerning the volume of data they can reasonably process. This information would be incorporated into rules under DBMS CHARACTERISTICS. Certain hardware also is limited or impractical in its ability to handle specified volumes of data, therefore additional rules on data volumes must be developed under HARDWARE.

The first primary area of influence, DBMS CHARACTERISTICS, relates to the DBMS' being evaluated and to the specific limitations/capabilities of each. The set of DBMS' that DBMSCON is currently being developed around is shown below.

<table>
<thead>
<tr>
<th>PRODUCT</th>
<th>VENDOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADABAS</td>
<td>Software AG</td>
</tr>
<tr>
<td>CREATEBASE</td>
<td>IDX</td>
</tr>
<tr>
<td>DB1012C</td>
<td>Teradata</td>
</tr>
<tr>
<td>DB/2</td>
<td>International Business Machines</td>
</tr>
<tr>
<td>EMS 170</td>
<td>Control Data</td>
</tr>
<tr>
<td>EASYTRIEVE</td>
<td>Pansophic</td>
</tr>
<tr>
<td>FOCUS</td>
<td>Information Builders</td>
</tr>
<tr>
<td>IDMS</td>
<td>Cullinet</td>
</tr>
<tr>
<td>IDM 500</td>
<td>Britton Lee</td>
</tr>
<tr>
<td>IMS</td>
<td>International Business Machines</td>
</tr>
<tr>
<td>INFETCH</td>
<td>Magna Systems</td>
</tr>
<tr>
<td>INGRES</td>
<td>Relational Technology</td>
</tr>
<tr>
<td>INQUIRE</td>
<td>Infodata</td>
</tr>
<tr>
<td>MODEL 204</td>
<td>Computer Corporation of America</td>
</tr>
<tr>
<td>NOMAD</td>
<td>D &amp; B Computing Services</td>
</tr>
<tr>
<td>ORACLE</td>
<td>Oracle</td>
</tr>
<tr>
<td>RAMIS</td>
<td>Mathematica</td>
</tr>
<tr>
<td>RIM</td>
<td>Boeing</td>
</tr>
<tr>
<td>SAS</td>
<td>SAS Institute</td>
</tr>
<tr>
<td>SIR</td>
<td>SIR</td>
</tr>
<tr>
<td>SQL/DS</td>
<td>International Business Machines</td>
</tr>
<tr>
<td>SYSTEM 2000</td>
<td>SAS Institute</td>
</tr>
<tr>
<td>TOTAL</td>
<td>CINCOM</td>
</tr>
</tbody>
</table>

Not all of these packages are full-fledged DBMS', nor should they be. In many cases there is little need for the full capabilities of a DBMS. There is little value in recommending a package such as IMS if one such as EASYTRIEVE will suffice. This is taken into account as each of the above packages is evaluated. The initial set of DBMS' selected to be a part of the knowledge base of DBMSCON was established based on the availability of a specific DBMS on Boeing computers, knowledgeable personnel within The Boeing Company, or its popularity in the DBMS marketplace.

With each of these DBMS' come limitations, both theoretical and practical, that influence selecting a DBMS. Limitations in security, data volume, number of users, level of implementation (i.e., implementation at a corporate or personal level) represent critical areas in selecting the right DBMS. Special capabilities and modes of processing also need to be reviewed for each DBMS. These cover things such as index processing, dynamic where-clause processing, interfaces to other DBMS', updating techniques, and special features in the query language(s).

Initially there were thoughts of incorporating detailed features of the query languages for each of the DBMS'. It was determined not to do so for several reasons: 1) the query language(s) of a DBMS can provide basic reporting functions and therefore do not have a major influence in the selection of a DBMS, 2) the special capabilities/limitations of a query language can be more easily identified and incorporated into a rule, 3) the detailed features of individual DBMS' change too frequently to attempt to maintain the information in an expert system, and 4) the primary criteria for selecting a DBMS should rarely be its detailed query capability, but rather its ability to manage data.
The flow time required to develop and implement a database may be essential from two aspects: 1) the learning curve for a DBMS, and 2) the number of staff personnel available to develop/maintain the data base applications and the DBMS. The learning curve for DBMSs' varies greatly from a basic user perspective as well as from the level of technical sophistication necessary to provide ongoing support. The personnel required to support the data base environment during and after development could be significant, and therefore must be understood before DBMS selection can be valid.

The second primary area of influence in selecting a DBMS is the DATABASE STRUCTURE. A knowledge of the data characteristics and data structure must be acquired. DBMSCON reviews the level of the requirements and logical design already completed. It attempts to identify the logical, intrinsic structure of the data as sequential (strictly one-to-one relationship), hierarchical (one-to-many), or network (many-to-many). This is done in part by attempting to identify the number of data fields (elements) and relations defined.

DBMSCON attempts to determine the physical database structure (sequential, hierarchical, network, inverted, or relational) that will be necessary, independent of a DBMS. DBMSCON does not eliminate any database structure that will not handle a specific data model but will downgrade them. A question asking the user for the data model needed would generally get a highly subjective response and therefore is not posed to the user directly. Two of the areas that need to be considered in order to resolve the database structure are the dynamics of the database environment and the usage characteristics.

The dynamics of the database environment being designed are primarily based on the continuing need for the database structure to change, and on the data volumes fluctuating over time. Each database structure has certain inherent characteristics that affect its ability to handle change. The ability for the database to change and the need for it to change must be evaluated closely. The extent to which data values change or new information loaded (including the mode in which it will change) also impacts the database structure chosen. This may even determine that information not be placed into a DBMS.

The usage characteristics indicate specific capabilities that are required of a DBMS and how users will interface with the database. DBMSCON attempts to determine the database and processing environment, and the type of searching that will be necessary to retrieve the data. Lines of questioning attempt to ascertain processing modes (sequential, random, etc.) access modes (batch/interactive), and application development (program language interfaces, areas of application development, etc.)

The HARDWARE environment must be determined. This is the third primary area of influence. Hardware constraints are frequently well defined. Most companies are not willing to change computer vendors, or even computers, just to satisfy their data base needs. Therefore, the evaluator may specify the hardware environment or have DBMSCON identify a suitable hardware environment. Currently, DBMSCON limits the realm of its hardware to CDC, Cray, DEC (VAX), and IBM. Even so, there are many factors to be evaluated in selecting the best hardware environment. Some of the key factors in selecting the hardware are data volume, number of users, level of implementation, and performance. DBMSCON evaluates these factors individually and collectively.

SUMMARY

Given sufficient time and money any application can be made to fit into any DBMS. The successful implementation of an information system merely by converting it to use a DBMS is, at best, a poor yardstick. Traditionally, the selection of a DBMS has been heavily influenced by two factors: 1) vendor solicitations and 2) knowledge (turning into a prejudice) of a specific DBMS. Selecting a DBMS using these factors can be costly, without anyone realizing it. Excessive resources, computing or human, can heavily negate benefits even though a "new technology" was finally used. DBMSCON helps to reduce prejudices and attempts to evaluate DBMS in an impartial manner.

DBMSCON is not being developed to specify the single, best DBMS. It is being developed to reduce the number of DBMS to consider in final, detailed evaluations. It will not reduce the effort required for requirements or logical design. It is intended to reduce the set of DBMS to evaluate and improve the confidence in the final selection of a DBMS. The most highly recommended DBMS' should be scrutinized with care and respect. Not all of the knowledge required to select the proper DBMS can be placed in a set of rules. Politics, existing knowledge of a DBMS, and specific costs are also factors which are not incorporated (intentionally) into DBMSCON. The DBMS recommendations can only be as good as the information supplied to it. If the information supplied is inaccurate or general then so will be the recommendations.

FUTURE

Developing DBMSCON into a viable tool within The Boeing Company was not the primary objective. Even so, out of this has come the realization of a tool that would be valuable in the complex world of DBMS selection. Plans are being considered to continue the development of DBMSCON into a comprehensive, reliable expert system.

There is a desire to expand the knowledge base of DBMSCON with additional DBMS' and Fourth
Generation Languages (4GL). This expansion and the effort required for DBMSCON to become a competent expert is expected to take approximately one year. Another long term goal is to expand the knowledge base of DBMSCON to also include micro-computer based DBMS and distributed computing system environments.

Mr. Van Rossum provides product and project support for Data Base Management Systems within the Information Management Group of Boeing Computer Services' Engineering Technology Applications Division. He has worked with database management systems for over ten years and has provided consultation to many companies and government organizations. For additional information on DBMSCON you may contact him via the following:

Boeing Computer Services
P.O. Box 58206
Seattle, WA 98124
ATTN: David Van Rossum, MS 96-23

or call:

(206)575-5276
Figure 1 -- Subject Areas Influence Recommendations of DBMS' within the Logical Hierarchy of DBMCON