# A Survey on Usage of SQL

*Hongjun Lu* *Hock Chuan Chan* *Kwok Kee Wei*
Department of Information Systems and Computer Science
National University of Singapore
10 Kent Ridge Crescent, Singapore 0511
Internet: {luhj,chanhc,weikk}@iscs.nus.sg

## Abstract

Relational database systems have been on market for more than a decade. SQL has been accepted as the standard query language of relational systems. To further understand the usage of relational systems and relational query language SQL in industries, we conducted a survey recently that covers various aspects of the usage of SQL in industrial organizations. In this paper, we present those results that may interest DBMS researcher and developers, including the profiles of SQL users, the application areas where SQL is used, the usage of different features of SQL and difficulties encountered by SQL users.

## 1 Introduction

Nowadays, relational database management systems (RDBMS) [Codd 1970, Codd 1990, Date 1989] can be obtained for various hardware platforms at the price most organizations can afford. SQL, one of the relational query languages, has been accepted as the standard industrial database query language [SQL 1987-1992]. Along with the evolution process of relational DBMS and its query languages, a number of studies have been conducted and reported [Codd 1988, Date 1986, Greenblat & Waxman 1978, Reisner 1977, Reisner 1981, Smelcer 1989, Welty & Stemple 1981, Welty 1985]. Most of the studies are laboratory studies. To further understand the usage of SQL by industries and hopefully find some useful implications to DBMS researchers, developers, educators, data processing managers, etc.. we conducted a survey study in Singapore. Singapore, as a developing country, the usage of DBMS may not be as popular as some other western countries. However, as the Singapore government has been especially promoting the IT technology, the number of users of relational DBMS and SQL has greatly increased. As such, the results obtained may be of general interests.

The general purpose of the study is to gain a broad view of the usage and acceptance of SQL in Singapore's IT industry. In particular, the objectives of the study include

- To develop a profile of the major SQL users and applications;
- To determine the extent to which SQL features are being utilized;
- To assess problems and difficulties that SQL users encounter when writing SQL queries;
- To identify related areas that can be of assistance in the development of more practical and effective training courses for current and potential SQL users;
- To identify ways of improving the use of DBMSs in the business community, such as introducing new technology, e.g. object-oriented systems.

Mail questionnaire survey was chosen to be the primary data collection method. A questionnaire consisting of 8 pages with 5 sections were designed for this study. Sections 1 and 2 gathered general information of the respondents, their organizations, and the purposes for the usage of data. Section 3 recorded information on the sources of respondents' SQL knowledge, the applications that the SQL queries are commonly used for, and comparison of SQL with other query languages. Section 1 to 3 contained multiple choice and fill-in-the-blank questions. Section 4 examined the extent of use of the common SQL features. It contained questions requiring the respondent to state the percentages of his queries that use the features. The last section captured information on the problems and difficulties in using SQL, and rated SQL as a database query language. It contained questions with 7-point scales. A final question allowed respondent to express additional views.

A total of 700 questionnaires were mailed to 175 selected companies, whose name are obtained from difference sources, including the suppliers of major RDBMS (ORACLE, INGRES, INFORMIX, etc.) in Singapore. Telephone follow-ups were carried out for reminder as well as for data clarification purposes. Of these, 149 questionnaires were returned, giving a response rate of 21.3%. These came from 41 (23.4%) of the companies. All questionnaires were checked for completeness and a total of 136 valid questionnaires remained useful for data analysis.

In this short paper, we report part of the results obtained from the survey which may be of interest to the readers of SIGMOD RECORD. For a more detailed report, readers may refer to [Chan *et al.* 1993].

# 2 Survey results

## 2.1 The profile of respondents

Table 1 to 4 summarize the general information of respondents, including their job positions, the years of computer related experience, the hardware platforms they used and the sources where they got their SQL knowledge.

| Job Title | Percentage |
|---|---|
| Systems Analyst | 25.0 |
| Systems Programmer | 0.7 |
| Systems/Software Engineer | 8.1 |
| Programmer Analyst | 20.6 |
| Application Programmer | 5.9 |
| Database Administrator | 11.0 |
| Project Leader | 8.8 |
| Consultant | 5.1 |
| Others | 14.7 |

Table 1: The Job Classification of Respondents

In Table 1 the job titles of respondents are summarized. Though the job titles does not classify the user precisely, it can be seen that programmers are the major users of the language. Systems analysts, systems/software engineers, programmer analysts and application programmers constitute 60.3 % of the respondents. Other users include DBA's, project leaders and consultants. There are 14.7% of respondents categorized as others which may include the end-users of the database systems. Relatively speaking, the percentage of this group is low as database query languages are designed for both programmers and end-users who are non-programmers. One possible reason is that the questionnaries were not distributed widely to those end-users.

Respondents who made use of a mixture of mainframe, mini- and micro- computers forms the largest group, with the next largest group using mainly mini-computers, as shown in table 2.

| Mainframe | Mini | Micro | Mixed | others |
|---|---|---|---|---|
| 11.8 | 30.1 | 15.4 | 36.0 | 6.6 |

Table 2: Type of Computers used.

Table 3 shows that the majority of respondents (39.7%) had 3-7 years of computer experience and only 29.4% had more than 7 years of experience.

| Years of Experience | $< 1$ | 1-3 | 4-7 | $> 7$ |
|---|---|---|---|---|
| Percentage | 5.9 | 25.0 | 39.7 | 29.4 |

Table 3: Experience with Computers.

About 40% of the respondents were first exposed to SQL through tertiary education, while 27.2% gained their initial knowledge from SQL manuals. Courses provided another big source (30.1%) for the language. However, SQL manuals was the major source of knowledge (58.1%) in the course of the users' work. These are shown in table 4. [1]

| Source of SQL Knowledge | Initial | Further |
|---|---|---|
| Tertiary Education | 39.7 | 2.9 |
| Manuals | 27.2 | 58.1 |
| External Courses | 15.4 | 30.9 |
| In-house Traing Courses | 14.7 | 22.1 |
| Others | 2.9 | 3.7 |

Table 4: Sources of SQL knowledge

## 2.2 The application areas of SQL

The majority of the respondents came from computer companies (30.9%), with banking/finance/insurance firms (13.2%) taking second place, as shown in Table 5. Of the responding organizations, 61.8% are private firms with 39.7% local firms and 22.1% foreign-based multinationals. The remaining 38.2% are from the public sector.

[1] The percentages for the further sources of SQL knowledge is more than 100 because some respondents have multiple sources of knowledge.

As for application areas, personnel (36.8%), accounting (28.7%) and finance (25.7%) were found to be the applications that SQL queries were most commonly used for. These were closely followed by inventory (23.5%), billing (21.3%) and payroll (19.1%) applications. Among applications that were classified under the 'others' category were operations control, planning, shipping, student records, and examination time-table generation. The growing importance of data resource [Niederman *et al* 1991] is confirmed by the usage of SQL on a broad coverage of applications, and the substantial amount of ad-hoc queries.

| *Industry Types* | |
|---|---|
| Industry Type | Percentage |
| Computer Company | 30.9 |
| Banking/Finance/Insurance | 13.2 |
| Education/Training/R&D | 10.3 |
| Transport/Communication | 8.1 |
| Manufacturing | 5.9 |
| Construction/Engineering | 0.7 |
| Others | 30.9 |
| *Application Areas* | |
| Business | Percentage |
| Personnel | 36.8 |
| Accounting | 28.7 |
| Finance | 25.7 |
| Inventory | 23.5 |
| Billing | 21.3 |
| Payroll | 19.1 |
| Costing | 16.9 |
| Marketing | 14.0 |
| Budgeting | 11.8 |
| Purchasing | 10.3 |
| Material control | 8.8 |
| Sales | 8.1 |
| Retailing | 5.1 |
| Others | 39.7 |

Table 5: Application areas of SQL queries

## 2.3 The complexity of SQL queries formed

Information on the type of queries written by SQL users is shown in Table 6 About half the queries are simple, using only 1 or 2 relations, which means without any joins, or at most a single join. However, a substantial 20.6% are complex queries that involve 5 or more relations.

| *No Of Rels, Attrs, Preds in a Query* | | | | |
|---|---|---|---|---|
| | 1-2 | 3-4 | 5-6 | > 6 |
| Relations | 51.8 | 27.6 | 11.8 | 8.8 |
| Attributes | 21.4 | 23.9 | 22.3 | 32.4 |
| Predicates | 38.8 | 33.5 | 16.1 | 11.6 |

Table 6: The complexity of queries

About half the queries (45.3%) involve only a few attributes (4 or less). But a substantial portion (32.4%) involve more than 6 attributes. About a third (38.8%) involve only 1 or 2 conditions, and slightly more than a quarter (27.7%) have 5 or more conditions, as shown in Table 6. These statistics are presented in a different way in Figure 1. The following classification is made to distinguish low, moderate and high complexity queries: low - 1 to 2 relations, 1 to 4 attributes or 1 to 2 conditions; moderate - 3 to 4 relations, 5 to 6 attributes or 3 to 4 conditions; high - 5 or more relations, more than 6 attributes, or 5 or more conditions. The figure shows that SQL usage is well spread from the simple to the very complex. Although SQL was originally meant for simple queries by end-users, it is now used by programmers for very complex queries. This is also confirmed by the results presented in the following subsection.

Different from some other relational language such as QUEL, SQL supports nested queries. Though nested queries are preferred by some users, deep levels of nesting often make the meaning of query obscure, especially when the inner subqueries involve variables in the outer subqueries [Codd 1990]. It is found that very small number of queries are nested more than three levels as shown in Table 7.

| No of levels | 1 | 2 | 3 | 4 | > 4 |
|---|---|---|---|---|---|
| percentage | 21.1. | 12.1. | 10.6 | 3.7 | 2.4 |

Table 7: Nested queries

## 2.4 Usage of various SQL features

The basic SELECT statement has three clauses, SELECT-clause, FROM-clause, and WHERE-clause. SQL also provides grouping and sorting facilities, the GROUP BY, ORDER BY and HAVING clauses. Queries using these features were considered hard. [Welty & Stemple 1981]. In the survey, it is found that GROUP BY and ORDER BY are in fact two heavily used features. The respondents indicated that about 43.1% and 57.6% queries they formed involving

Figure 1: Percentages of queries by different complexity measures

these two features as shown in Table 8. As expected, HAVING clause is less frequently used – most students we taught have had difficulties to distinguish WHERE conditions from HAVING conditions and tend to use WHERE clauses if possible.

| GROUP BY | HAVING | ORDER BY |
|---|---|---|
| 43.1% | 20.8% | 57.6% |

Table 8: Grouping and sorting

Set operations of SQL, such as UNION and INTERSECT, are also thought to be difficult to use. [Welty & Stemple 1981] Compared to grouping and sorting facilities, the set operations are less used, as shown in Table 9.

| INTERSECT | UNION | MINUS |
|---|---|---|
| 25.6% | 24.4% | 10.4% |

Table 9: Set operations

SQL aggregate functions, SUM, COUNT, MAX, MIN, AVG, are rather useful features, especially in generating reports. However, proper use of the aggregate function is not as easy as simple select statement. Two most commonly used functions are SUM (40.1%) and COUNT (37.6%) (Table 10).

The usage of other features is also surveyed. For example, the percentage of queries using the existence quantifier, EXIST or NOT EXIST is 25.2%. The overall results give the impression that most SQL features are used rather often in real applications.

| SUM | COUNT | MAX | MIN | AVG |
|---|---|---|---|---|
| 40.1% | 37.6% | 19.9% | 18.1% | 17.6% |

Table 10: Aggregate functions

### 2.5 Difficulties encountered and actions

SQL studies, over the years, have shown that the percentage of incorrect SQL queries have varied from 12% to an unacceptable rate of 75% [Reisner 1977, Greenblat & Waxman 1978, Welty 1985]. These studies made a broad coverage of the possible SQL errors, both syntactic and logical. From the survey results shown in Table 11, the majority (82%) of the respondents usually make more than one before obtaining the desired results. Only 4% had to try more than five times while an elite group of 18% would get the results at their first attempt.

| 1 time | 2-4 times | 5-8 times |
|---|---|---|
| 18% | 78% | 4% |

Table 11: Number of attempts to obtain expected query result

Basically, making two to four attempts did not seem to deter the users very much, as demonstrated by the overall satisfaction for using SQL. Nevertheless, respondents did indicate problems which they encountered while using SQL. These include:

- comprehension difficulty

- complex queries are difficult to analyse, especially by another person
- "nested maze" is quite confusing. This confirms one of the theoretical flaws of SQL - not well defined semantics for nesting (Codd 1990).
- multiple joins of many tables can lead to uncertainty of the query accuracy
- logical errors are harder to detect, as compared to 3GLs

- formulation problem
  - joins are difficult for end-users
  - too many aggregate functions in a single query have led to problems
  - use of wrong field and name definition
  - unable to format the output as desired
  - variables used with wrong variable types, especially for embedded SQL
- performance
  - response is slow when system does not select the best path to access tables.
  - database contention occurs by simultaneous accesses
  - a query may need to be broken into smaller queries to speed up processing time. This requires more temporary space.
- unclear error message sometimes give wrong impressions

When users encounter problems with SQL, the majority (68%) refer to manual. This also confirms the finding that manuals form a substantial secondary source of SQL knowledge. A substantial 24% prefer to seek the assistance of colleagues or superiors. Only a minority, 2%, attempt querying with other languages, while 6% will try other means, one of which was to try till I get it right. to SQL manuals (Table 12).

| Actions Taken | Percentage |
|---|---|
| Refer to manuals | 68 |
| Get colleagues' help | 20 |
| Get superiors' help | 4 |
| Use other languages | 2 |
| Others | 6 |

Table 12: Actions taken when encountering SQL problems

# 3 Discussion

Literature materials that compare SQL with other database query languages are not hard to find. However, these are mostly done in the laboratories with student users. Limited research has been done to study the language in terms of drawing up a classification of the SQL users, the SQL features commonly used, and the problems SQL users encountered. In this paper, we present part of the results obtained from a recent survey conducted in Singapore. The participants of the survey are SQL users in the "real" world. Although the sample size of the study is rather small and may not be random, the findings may be still of some interests to researchers, developers and educators.

The results of survey indicated that SQL and relational DBMS are used in various business areas. Although most SQL users do not have many years of computer related experience, various features of SQL have been widely used and queries formed are rather complex in terms of the number of relations, attributes, predicates involved and the number of levels of nesting. On other hand, most users need to try a few times to get the answers they expected. Reference manuals and training courses seem to be very important to make the use of SQL more popular and effective. We believe that, from the statistics gathered from this survey, individuals and industries who are using SQL and relational DBMS, vendors of relational systems, academic researchers may be able to derive some lessons.

As mentioned previously, the coverage of this survey is not large enough. More comprehensive survey is needed to confirm the findings. It is hoped that the future survey can cover more industry and business organizations, reach more users in different areas and with different experience. Other features of the SQL language, such as data definition facilities, updating facilities, etc. should also be included to make the evaluation as complete as possible.

# Acknowledgment

The authors like to thank Liang Boon Yen for his contributions to the survey study.

# References

[Chan *et al.* 1993] Chan, H.C., Lu, H., and Wei, K.K., "*A Survey of SQL Usage,* " Technical Publications, Department of Information Systems and

Computer Sciences, National University of Singapore, June 1993.

[Codd 1970] Codd, E.F., "A Relational Model of Data for Large Shared Data Banks," *Communication of the ACM* Vol. 13, No.6, June 1970.

[Codd 1982] Codd, E.F., "Relational Database: A Practical Foundation for Productivity," *Communication of the ACM* , Vol. 25, No.2 New York: Association for Computing Machinery, Inc., Feb 1982.

[Codd 1988] Codd, E.F., "Fatal Flaws in SQL," *Datamation* , Aug 1988, 45-48, and *Datamation* , Sept 1988, 71-74..

[Codd 1990] Codd, E.F. (1990). *The Relational Model for Database Management,* Version 2, Addison-Wesley Publishing Company, 1990.

[Date 1986] Date, C.J., "A Critique of the SQL Database Language", in *Relational Database: Selected Writings* , Addison-Wesley Publishing Company, 1986.

[Date 1987] Date, C.J., *A Guide to the SQL Standard,* Addison-Wesley Publishing Company, 1987.

[Date 1989] Date, C.J., *Relational Database Writings 1985-1989,* Addison-Wesley Publishing Company, 1989.

[Dieckmann 1981] Dieckmann, E.M., "Three Relational DBMS, " *Datamation,* Sept 1981, 137-148.

[Greenblat & Waxman 1978] Greenblatt, D. and Waxman, J., "A Study of Three Database Query Languages," in *Databases: Improving Usability and Responsiveness,* Ed. B. Shneiderman, New York: Academic Press, 1978, 77-97.

[Greene *et al.* 1989] Greene, S.L., Devlin, S.J., Cannata, P.E., and Gomez, L.M., "No IFs, ANDs, or ORs: A Study of Database Querying, " *Database Querying,* 1989, 303-325.

[SQL 1987-1992] International Organization for Standardization (ISO), *Database Language SQL,* Document ISO/IEC 9075:1987, 9075:1989, 9075:1992.

[Lusardi 1988] Lusardi, F., *The Database Experts' Guide to SQL,* Intertext Publications/ Multiscience Press, Inc., 1988.

[Niederman *et al* 1991] Niederman, F., Brancheau, J.C., Wetherbe, J.C., "Information Systems Management Issues for the 1990s," *MIS Quarterly,* Dec 1991, 475-500.

[Reisner 1977] Reisner, P., "Use of psychological experimentation as an aid to development of a query language". IEEE Transactions on Software Engineering, SE-3, 1977, 218-229.

[Reisner 1981] Reisner, P., "Human Factors Studies of Database Query Languages : A Survey and Assessment," *Computing Surveys,* Vol. 13, No. 1, March 1981, 13-31.

[Samet 1981] Samet, P.A., *Query Languages : A unified approach,* Heyden and Son Ltd., Cambridge, 1981.

[Smelcer 1989] Smelcer, J.B. , *Understanding User Errors in Database Query,* Doctoral thesis, University of Michigan, Ann Arbor, USA, 1989.

[Welty & Stemple 1981] Welty, C., and Stemple, D.W. , "Human Factors Comparison of a Procedural and a Nonprocedural Query Language," *ACM Transactions on Database Systems,* Vol. 6, No. 4, 1981, 626-649.

[Welty 1985] Welty, C. "Correcting User Errors in SQL," *International Journal of Man-Machine Studies,* No. 22, 1985, 463-477.