

# A Spreadsheet Exercise for Teaching Data Volume and Usage Analysis in the Introductory Database Course

**Norman Pendegraft**  
**College of Business and Economics,**  
**University of Idaho, Moscow, ID, USA**  
[norman@uidaho.edu](mailto:norman@uidaho.edu)

## Abstract

It is argued that volume and usage analysis are useful topics in the introductory database course. Since time is limited and many students struggle with the issues, a simple spread sheet model is presented for helping students learn data volume and usage analysis. A simple example is presented, and student response to a related exercise is reported.

**Keywords:** database volume analysis, usage analysis

## Introduction

Data volume and usage analysis are interesting topics in database design or administration courses. Since the introductory course may be the only database course taken by many students, if issues of physical design are to be introduced, it must be in this course. Issues include estimating the size of the database and examining likely usage patterns to look for opportunities for physical design improvements like clustering. The author finds the discussion useful in introducing matters of physical design and tuning. However, the typical introductory database course may not have much time for discussion of these issues. Tuning is only mentioned by one panelist in two separate discussions (Adams et.al., 2004, Robbert et.al, 2000). Robbert & Ricardo (2003) report that optimization is typically allocated less than 2 hours per semester. In short, while these issues may be interesting there is not a lot of time available for this discussion, hence the need to make efficient use of the time allocated.

Motivating discussion of volume analysis is easy: one simply asks, “About how big will this database be?” The question is readily perceived as interesting and important. Some motivation for usage analysis is provided by suggestions like the following from the Oracle Performance Tuning Guide (Oracle, 2008)

“Consider indexing keys used frequently in where clauses ... [or are] frequently used to join tables.” (14-3)

---

Material published as part of this publication, either on-line or in print, is copyrighted by the Informing Science Institute. Permission to make digital or paper copy of part or all of these works for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage AND that copies 1) bear this notice in full and 2) give the full citation on the first page. It is permissible to abstract these works so long as credit is given. To copy in all other cases or to republish or to post on a server or to redistribute to lists requires specific permission and payment of a fee. Contact [Publisher@InformingScience.org](mailto:Publisher@InformingScience.org) to request redistribution permission.

“Cluster tables that are accessed frequently ...[via] join[s]...

“Cluster master detail tables if you often select a mater record and then the corresponding detail record...

“Do not cluster tables if ...[you] modify the common values frequently ...[as this] takes longer ... because Oracle might need to migrate the modified row to another block.” (14-10).

## Spreadsheet Exercise

While such suggestions make perfect sense to an experienced database designer, they may appear arbitrary to a novice. Offering some theoretical discussion can illuminate the issue and make it clear that these “rules” are not arbitrary, but follow naturally from the way the data is actually stored. Usage analysis provided an opportunity to both raise the issue and also to shed light on strategies for physical design.

The computer science literature holds much serious analytical and empirical work examining physical design issues. Christodoulakis (1984) examined the implications of assumptions like constant number of records per block. Such assumptions often offer only bounds on cost. Clearly, if such a detailed analysis gives only bounds, then a brief unit in the introductory course can do little more. In particular detailed discussion of block sizing and chaining are not appropriate in the first course. Finkelstein and others (1988) developed an automated optimizer to allocate indexes. They also offer an excellent discussion of the economics of index configuration. Their system takes a set of SQL queries and frequencies as input. The usage analysis example described here, essentially provides a way to identify this set of queries and estimate their frequencies. While the resultant analysis is in no sense optimal, it does serve to illustrate the issue which is the objective of the exercise.

While the problems are not technically difficult, they prove to be, in the author’s experience, very challenging for many students, perhaps because they are not accustomed to “back of the envelope” estimations. The topic is not addressed in all texts, and in those that do, the discussion proves difficult for many students to grasp. Further, this analysis is not readily supported by many CASE tools.

A secondary opportunity presented by the discussion is introduction to the use of orders of magnitude for estimating database volumes. The use of orders of magnitude is a common method in the sciences, but one which seems largely unfamiliar to typical business majors.

Finally, the exercise requires that the students develop a spreadsheet to perform many of the calculations directly. It is a natural extension of a tool that they already know into analysis in a novel application, adding the additional advantage of emphasizing the spreadsheet as an analytical tool rather than a computational tool.

## Example

The entity relationship diagram (ERD) in Figure 1 illustrates a simple database design appropriate for the exercise. Additional facts regarding volumes and transactions are offered as well. Note that the data are given in the form of business transactions, not SQL statements, as that is the way that the analyst will encounter them. The ERD in Figure 2 uses the notation of Hoffer et.al. (2011) to represent the volume and transaction data on the ERD. In particular, the numbers in the entities represent the number of records estimated. The arrows represent query paths: the number indicates the rate and a “Q” represents a query while an “I” represents an insert on the database.

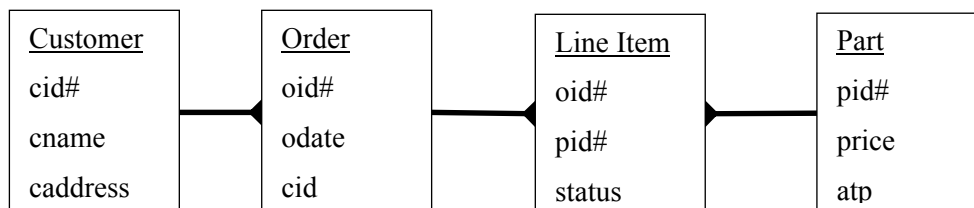


Figure 1: Example Database Design

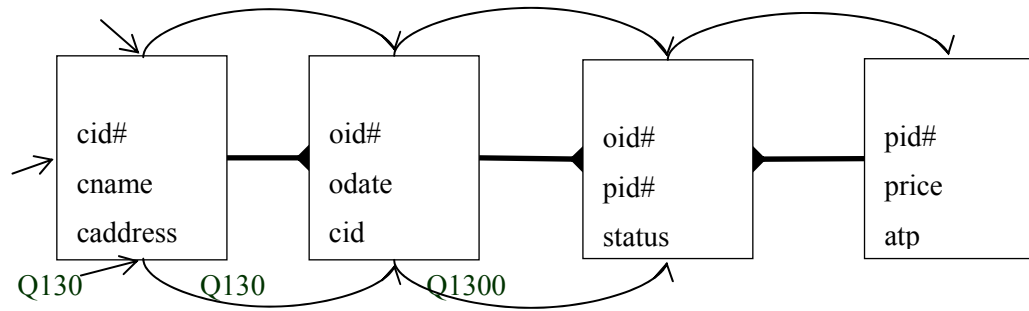


Figure 2. Volume and Usage Analysis

### Current Data Volumes

1. There are currently 90,000 customers.
2. The catalog has 6865 SKUs.
3. There are currently 750,000 orders pending.
4. Order averages 10 items.

### **Transactions on the Database**

1. Approximately 10 new customers place orders each hour.
2. Approximately 800 orders are place each hour.
3. Approximately 10 new items are added to the catalog each month.
4. About 4000 queries are received each day that require examining the catalog, but do not result in sales.
5. Approximately 200 customers call each hour to inquire about pending orders, requiring looking up all pending orders and item details.

### **Orders of Magnitude.**

During the volume analysis discussion in class, it is helpful to review the notion of order of magnitude (OOM). Unfortunately, for many students, this is an introduction, not a review. The idea of orders of magnitude is commonly used in the sciences when the accuracy of the data is limited, or when precise calculations are too time consuming. (See for example, Goldreich et.al., 1999).

The order of magnitude of a number is essentially the exponent when the number is written in scientific notation. For example, the order of magnitude of 1283 ( $1.283 * 10^3$ ). Usually, the "cutoff" is 3. That is if the mantissa is less than three, the number is rounded down, if above 3, then the number is rounded up. This stems from the fact that 3 is about  $10^{1/2}$ . Thus, in terms of exponents 3 is about "half" way between 1 and 10.

Using OOM has several advantages: first it is easier to estimate; it is easier to do the calculations; it does not convey an inappropriate sense of precision. The latter is especially important when one is estimating uncertain values. In the current context it also helps deal with the issue of database overhead. As noted above, the introductory course does not have time to address such matters in detail, but they can be subsumed into the model by adding a reasonable percentage to the overall size. In the end, it generally does not matter much because the issue is something like, "Will this fit on one (or a few) 100GB hard drive(s)?"

### Volume Analysis Spreadsheet

A simple spreadsheet tool was developed. The spreadsheet has two regions. One contains estimates of the number of records and record size and volume calculations for each table in the database. The second addresses the transactions executed on the database and relationships student assignment suggested that it was an improvement over the previously used method.

The spreadsheet is demonstrated in Figure 3 illustrate the volume analysis. Each table is listed, and for each table, then number of entries is estimates as is the approximate size of each tuple. In each row in the spreadsheet the product of these two numbers estimates the size of the table. The sum of those products gives an estimate for the size of the database. Note that this does not include the overhead imposed by the database itself. This idea needs to be included in the discussion, but in general, may be assumed to be small compared to the size of the data.

Note that both numbers are estimated as orders of magnitude. Note that the estimate is also an order of magnitude. The actual number is 1,041,865,000. Thus there is little difference between the “actual” number and the OOM estimate. One of the ideas that comes from this analysis is that the most important estimates are the number of customers and the size of each record. In particular note that even if the number of parts increases by a factor of 10, there is no significant impact on the overall size of the database.

	Number of Record	Record Size (B)	
customer	1.E+05	1.E+03	1.E+08
order	1.E+06	1.E+03	1.E+09
lineitem	3.E+06	1.E+00	3.E+06
part	1.E+05	2.E+01	2.E+06
			1.E+09

Figure 3: Volume Analysis

### Usage Analysis Spreadsheet

The purpose of usage analysis is to evaluate the nature of the “interactions” between the tables in the database. This might be used in order to help make decisions about denormalizing the design, creating clusters or indexes. The notation commonly used while graphical, presents problems in that the case tools we use do not support it, nor does it make the calculations easy.

The spreadsheet in Figure 4 illustrates the usage analysis. Each event in the list above is listed and is broken into the steps necessary for its execution. For each step the relationship is identified and the number of times the step must execute is calculated. For the first step in a transaction, the number of accesses to the table is simply the number of transactions. For subsequent tables, the number of hits is equal to the number of hits on the previous table in the transaction times the ratio between the number of records in the second table to each record in the first along this relationship.

Usage Analysis								
EID		Q/t	Event	Table2	rel	Table1	Rel Vol	Volume
1	new customer	10	insert	-		customer	1	10
2	order	1000	query	-		customer	1	1000
			insert	customer	places	order	1	1000
			insert	order	OcontainsLI	lineitem	10	10000
			query	lineitem	LcontainsP	part	1	10000
3	order query	300	query	-		customer	1	300
			query	customer	places	order	3	900
			query	order	OcontainsLI	lineitem	10	9000
4	new part	1	insert	-		part	1	1
5	catalog query	3000	query	-		part	1	3000

**Figure 4: Usage Analysis**

For example, transaction 3, new order query, requires finding the customer record, then pulling all pending orders for that customer, and then pulling all line items associated with those orders. There are 300 of these per hour. Each customer's average 10 orders and orders average 10 line items. Thus there are 300 hits per hour on the customer table followed by 100 hits on the order table followed by 1000 hits on the line item table.

The table is enhanced by highlighting those paths in the top half and top 10% of traversal frequency. These paths might be likely candidates for clustering. In the example, it might be desirable to cluster line item around orders since that path is the most heavily traversed. The spreadsheet could be further enhanced by condensing the table to delete duplicates. This would be easier to do by moving the data to a database.

## Results

This tool was introduced in fall 2011. Anecdotally, the student response was favorable. The class was asked to respond to a written assessment (approved by the institutional review board) regarding their spreadsheet experience as well as the topic itself. 14 usable assessments were completed from a class of 17 (82%). Because of the small numbers only simple statistics are reported.

The team based project included a volume / usage analysis and all teams performed adequately on the assignment. In addition, there was a volume analysis question on the final exam had an average of 6.9 out of 12 with 3 very low scores (3 or less) and 6 high scores (9 or better). It is entirely possible that the superior performance was due to the attention drawn to the topic by the assessment. It is also likely that those who did poorly on the exam were among those who did not work on the assignment. The responses indicate that at least 4 of those responding spent less than one hour on the assignment. Two of those admitted to not working on it all.

It seems clear from the data (summarized in Table 1) that knowledge of spreadsheets was not an issue. The average time spent was over 3 hours, so it is clear that most students spent considerable time on the assignment. All 14 agreed that using a spreadsheet to conduct the analysis made sense.

<b>Table 1: Data Summary</b>				
For each statement, indicate your level of agreement				
I am confident in my ability to do volume and usage analysis.				
Strongly disagree 0	Disagree 2	Neutral 3	Agree 8	Strongly Agree 1
I performed a volume analysis for the assignment				
Strongly disagree 1	Disagree 0	Neutral 1	Agree 9	Strongly Agree 4
I have a great deal of experience using a computer spread sheet.				
Strongly disagree 0	Disagree 0	Neutral 2	Agree 7	Strongly Agree 5
Using a spread sheet tool to do volume and usage analysis saves me time in the long run.				
Strongly disagree 0	Disagree 0	Neutral 0	Agree 9	Strongly Agree 5
How much time did you spend developing the spreadsheet? mean = 3 hours				

Thus, the author concludes that the assignment was successful. How much can be attributed directly to the spreadsheet is unknowable. Clearly the time spent on the assignment itself was an important part of the success achieved. However, since the exercise provides one more opportunity to display the power of a spreadsheet as a tool of analysis rather than simple computation, the author will use this approach again and hopes that others may find the report useful in developing their own assignments.

## Acknowledgements

The author thanks the editor and several anonymous referees for their very helpful comments.

## References

- Adams, E. S., Goelman, D., Granger, M., & Ricardo, C. (2004). Managing the introductory database course: What goes in and what comes out? *ACM SIGCSE 04*. Norfolk Va. Retrieved 20 Feb. 2012 from [http://delivery.acm.org/10.1145/980000/971467/p497-ad-ams.pdf?ip=129.101.96.123&acc=ACTIVE%20SERVICE&CFID=86290938&CFTOKEN=51123981&\\_acm\\_=1329930663\\_69d65d4fe21e4ce609257047ea93974b](http://delivery.acm.org/10.1145/980000/971467/p497-ad-ams.pdf?ip=129.101.96.123&acc=ACTIVE%20SERVICE&CFID=86290938&CFTOKEN=51123981&_acm_=1329930663_69d65d4fe21e4ce609257047ea93974b)
- Goldreich, P., Mahajan, S., & Phinney, S. (1999). *Order-of-magnitude physics*. University of Cambridge. Retrieved draft 20 Feb 2012 at <http://www.inference.phy.cam.ac.uk/sanjoy/oom/book-letter.pdf>
- Hoffer, J. A., Ramesh, V., & Topi, H. (2011). *Modern database analysis* (10th ed.). Boston: Prentice Hall.
- Oracle. (2008). *Oracle database performance tuning guide, 11g Release 1 (11.1)B28274-02*. Available at [http://docs.oracle.com/cd/B28359\\_01/server.111/b28274/title.htm](http://docs.oracle.com/cd/B28359_01/server.111/b28274/title.htm)
- Robbert, M. A., & Ricardo, C. M. (2003). Trends in the evolution of the database curriculum. *Proceedings of the 8th annual conference on Innovation and technology in computer science education (ITiCSE)*

'03), David Finkel (Ed.). ACM, New York, NY, USA, 139-143. DOI=10.1145/961511.961551  
<http://doi.acm.org/10.1145/961511.961551>

Robbert, M. A., Wang, M., Guimaraes, M., & Myers, M. E. (2000). The database course: What must be taught. [SIGCSE 2000](#): 403-404

## Biography

**Norman Pendegraft** is Professor of management Information Systems in the College of Business and Economics at the University of Idaho. He teaches database design and telecommunications management. His research interests include modeling and simulation of information security economics and information systems pedagogy.