Summary Data: The New Frontier

by W. H. Inmon

Inmon Home Page

With the acceptance of data warehousing as an accepted architecture come new approaches and techniques for the management of data. The first and most obvious issue of data management that comes with data warehouse is that of the challenge of managing large amounts of data. Hardware approaches, design approaches, and software approaches all are used to address the volumes of data that come with data warehouse. But managing massive amounts of data is not the only challenge that comes with data warehousing. Data warehousing also entails the gathering and the management of summary data. And summary data brings its own set of properties and peculiarities.

SUMMARY DATA

The world of data processing has not easily accepted or dealt with summary data. In the early days of structured programming and design it was suggested that no summary data ever be stored. Instead, the structured theoreticians told us that only detailed data should be stored. The notion was that if there was detailed data on hand, that summary data could always be calculated from detailed data. By doing dynamic calculations, the summary data was sure to be up to date. From that crude understanding of summary data as envisioned by the early theoreticians, the world of data warehousing has fully embraced summary data, recognizing summary data as an important and integral part of the decision making landscape.

What are some of the important considerations of the management of summary data? The first (and perhaps the most basic) consideration of summary data in any form is that summary data does not exist in an abstract form. Summary data ALWAYS has associated with it - explicitly or implicitly - some form of calculation. Stated differently, in order for summary data to exist, the summary data must have gone through some form of a calculation process. The calculation process may be simple or complex, but there is no such thing as summary data without a corresponding calculation. The calculation process generally is comprised of:

- what data was included in the summarization process,
- what data was excluded from the summary process,
- what formula was used in the calculation,
- when was the summarization made, and so forth.

And there may be other aspects of summarization that are of interest, such as who made the summarization, where was the summarization made, etc.

When looking at the summary data then, the DSS analyst must ALWAYS qualify the summary data with the process that created the summary. Unlike detailed data that merely requires a definition, summary data carries with it a different kind of baggage. The definition of summary data and the calculation used to produce the summary data is required as the barebones minimum for the understanding of summary data. Separating summary data from its essential calculation is a mistake.

DIFFERENT TYPES OF SUMMARY DATA

But the marriage of summary data with a calculation process is not the only unique aspect of summary data. Another interesting aspect of summary data is that summary data has many forms, and the different forms of summary data make all the difference in the understanding, usage, and management of the data. A convenient way of thinking about the different forms of summary data is to envision summary data existing in the form of a rainbow.

DYNAMIC SUMMARY DATA

The rainbow of summary data can be classified in terms of degrees of dynamic summary data and static summary data. Dynamic summary data is summary data whose accuracy depends on the moment in time the summarization is made. For example, suppose a large bank wants to know the collective balance of a large corporate customer such as IBM. IBM may have hundreds of accounts with the bank. While the bank certainly wants each account that IBM has to be managed properly, the bank also would like to have an up to the second accounting of the status of all the accounts IBM has in a single place. At 10:16 am on Monday the bank does a poll of all the accounts that IBM has and finds that IBM has a collective balance of $10,966,114 at that moment in time. The accuracy of the summarization is relevant only to the moment in time that the calculation is made. At 3:42 pm on Monday afternoon another corporate accounting for IBM is made and at this point in time IBM has $11,107,118 on hand. Such calculations are made up of dynamic summary data.

STATIC SUMMARY DATA

The other side of the rainbow from dynamic summary data is static summary data. Static summary data is data that is repeatable in its calculation. The calculation will be the same regardless of when the calculation is made. Suppose that a DSS analyst wishes to calculate corporate expenses for the previous quarter. The DSS analyst calculates that there were expenses of $1,107,115 for the quarter on the 15th of the month. Now suppose that - for whatever reason - the DSS analyst recalculates the quarterly expenses a week later. The DSS analyst will calculate a value of $1,107,115. If any other number is reached either the wrong data was used or the calculation was made incorrectly. Static summary data yields the same results every time the calculation is made, regardless of the moment in time when the calculation is done.

Because static data is able to be recalculated, it is ideal for inclusion in the data warehouse. There is no point in having to recalculate the static summary data repeatedly. But dynamic summary data may actually be dangerous to store in a data warehouse. If dynamic summary data is stored, then a decision made as of some other time than when the calculation was made may will be a very misleading decision. Using the dynamic summary made at 9:12 am in the morning at 5:49 pm in the afternoon is not a proper thing to do at all.

Because of these essential differences, dynamic summary data is commonplace in the ODS environment and static summary data is commonplace in the data warehouse environment.

LIGHTLY SUMMARIZED

But the types of summary data do not end with dynamic an static summaries. Another dimension of the rainbow of summarization is that of data either being lightly summarized or highly summarized. Data that is lightly summarized is data with a low level of granularity. For example, suppose a sales file is to be summarized. Summarizing sales by day by zip code may well produce a low level of summarization. Suppose there are 10,000 sales to be summarized and that summarizing sales by day by zip code produces a summary file with 7,500 entries. The 10,000 records have been reduced by only a small fraction. The reduction is indicative of very lightly summarized data.

HIGHLY SUMMARIZED DATA

Now suppose that the same detailed sales file is summarized by month by state. The summarization of 10,000 records produces 550 records. The summarization by month by state is at a much higher level of summarization than summarization by day by zip code.

As a rule, the higher the level of summarization, the more interest the summarization is to management. Conversely, the lower the level of summarization, the greater the chance that the summarization will be of interest to the clerical level. Conversely, the higher the level of summarization, the less flexible the data, and the lower the level of summarization, the more flexible the data.

A grid with different parameters can be created. One set of parameters are dynamic summary data and static summary data, and another set of parameters are highly summarized data and lightly summarized data.

Figure 1 shows that one set of parameters are dynamic summary data and static summary data, and another set of parameters are highly summarized data and lightly summarized data. These parameters can be arranged into a grid.

There tends to be many forms of lightly summarized data - both dynamic and static. But the more highly data becomes summarized, the more static it becomes.

COMMUNITIES OF USAGE

Another way to perceive the spectrum of data is through the communities who use the different forms of data. Management tends to look at highly summarized data and the clerical community tends to look at lightly summarized data. Long term decisions tend to be made from static summary data and short term decisions tend to be made from both static and dynamic summary data. There are many clerical decisions that are made using both dynamic and static summary data. And there are many management decisions made using static data. But there are relatively few management level decisions made using dynamic data.