Capacity Planning

Any data warehouse solution will grow over time, sometimes quite dramatically. It is essential that the components of the solution (hardware, software, and database) are capable of supporting the extended sizes without unacceptable performance loss, or growth of the load window to a point where it affects the use of the system.

The capacity planner is interested in the access to the data warehouse, the DBMS capabilities and efficiencies, the indexing of the data warehouse, and the efficiency and operations of storage. Each of these aspects plays a large role in the throughput and operations of the data warehouse.

Calculating storage requirement

The calculations for space are almost always done exclusively for the current detailed data in the data warehouse. The reason why the other levels of data are not included in this analysis is that:

- They consume much less storage than the current detailed level of data, and
- They are much harder to identify.

The calculations for disk storage are very straightforward. Figure 1 shows the elements of calculation.

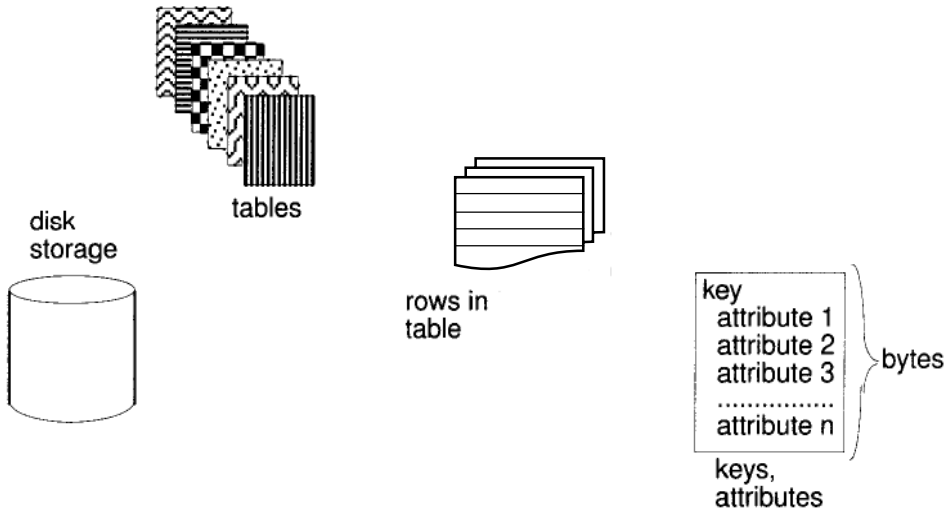

Figure 1 *Estimating disk storage requirements for the data warehouse*

To calculate disk storage, first the tables that will be in the current detailed level of the data warehouse are identified. In practice, only the very largest tables need be identified. Usually there are a finite number of those tables in even the most complex of environments.

Once the tables are identified, the next calculation is how many rows will there be in each table. Of course, the answer to this question depends directly on the granularity of data found in the data warehouse. The lower the level of detail, the more the number of rows. In some cases the number of rows can be calculated quite accurately. Where there is a historical record to rely upon, this number is calculated. For example, where the data warehouse will contain the number of phone calls made by a phone company's customers and where the business is not changing dramatically, this calculation can be made. But in other cases it is not so easy to estimate the number of occurrences of data.

After the number of rows are discovered, the next step is to calculate the size of each row. This is done by estimating the contents of each row - the keys and the attributes.

Once the contents of the row are taken into consideration, the indexes that are needed are factored in.

In practice, very little if any free space is left in the data warehouse because data is not updated in the warehouse. In most circumstances, any free space in a data warehouse is wasted.

The total disk requirements then are calculated by adding all the requirements mentioned.

Processor Requirement

In order to make sense of the estimation of the processor requirements for the data warehouse, the work passing through the data warehouse processor must be divided into one of three categories - background processing, predictable DSS processing, and unpredictable DSS processing.

1. Background Processing

Background processing is that processing that is done on a predictable, (usually) batch basis. Typical of background processing is extract processing, data warehouse loads, monitors, sorts/merges, restructuring, index creations, etc. Background processing is that utilitarian

processing necessary to the data warehouse but not directly associated with a query or an analysis of data warehouse data.

Background processing can be run at off peak times and can be spread evenly throughout the day. There is seldom much of a time constraint for background processing.

2. Predictable DSS processing

Predictable DSS processing is that processing that is regularly done, usually on a query or transaction basis.

The parameters of interest for the data warehouse designer (for both the background processing and the predictable DSS processing) are:

- The number of times the process will be run,
- The number of I/Os the process will use,
- Whether there is an arrival peak to the processing,
- The expected response time.

These metrics can be arrived at by examining the pattern of calls made to the DBMS and the interaction with data managed under the DBMS.

3. Unpredictable DSS processing

The third category of process of interest to the data warehouse capacity planner is that of the unpredictable DSS analysis. The unpredictable process by its very nature is much less manageable than either background processing or predictable DSS processing. Even the simplest of the task such as addition of variables that have not been regularly added before could be unpredictable DSS processing task. However, certain characteristics about the unpredictable process can be projected (even for the worst behaving process.) For the unpredictable processes, the:

- Expected response time (in minutes, hours, or days) can be outlined,
- Total amount of I/O can be predicted, and
- Whether the system can be paused temporarily during the running of the request can be projected.

Once the workload of the data warehouse has been broken into these categories, the estimate of processor resources is prepared to continue.

After the peak hour for I/O requirement is calculated (for example office time from 9:00 am to 5:00 pm), the next step is to determine what the hour-by-hour requirements are.

After the hourly calculations are done, the next step is to identify the "high water mark." The high water mark is that hour of the day when the most demands will be made of the machine.

After the high water mark requirements are identified, the next requirement is to scope out the requirements for the largest unpredictable request. The largest unpredictable request must be parameterized by:

- How many total I/Os will be required,
- The expected response time, and
- Whether other processing may or may not be paused.

If no temporary pause is allowed, then the largest unpredictable request is simply added as another request. If some of the workload (for instance, the predictable DSS processing) can be paused, then the largest unpredictable request is added to the portion of the workload that cannot be paused. If all of the workload can be paused, then the unpredictable largest request is not added to anything.

The analyst then selects the larger of the two - the unpredictable largest request with pause (if pause is allowed) or the unpredictable largest request added to the portion of the workload that cannot be paused. The maximum of these numbers then becomes the maximum processing capability needed for the system.

Approaches for DW capacity planning

There are broadly three approaches for capacity planning for the data warehouse, DSS environment. They are:

i. The Analytical Approach

The first approach to capacity planning is the analytical approach. The analytical approach is one in which the capacity planner attempts to calculate and/or predict capacity needs before the equipment is purchased. In the analytical approach the analyst attempts to quantify such things as:

- how many customers will be in the warehouse;
- at what rate will the customers grow;
- how many transactions will be in the warehouse;
- at what rate will the transactions grow;
- what other data will be in the warehouse;
- at what rate will the other data grow;
- what is the proper level of granularity for data in the warehouse;
- can the level of granularity be changed if needed;
- what amount of history is needed in the warehouse;
- will the user decide to add more history than anticipated? etc

Each of these interrelated questions must be answered in order for the analyst to determine how much data will be in the warehouse. But volumes of data are only one aspect of capacity planning. The other side of capacity planning in the data warehouse, DSS environment is that of workload projection.

ii. The Calibrated Extrapolation Approach

The calibrated extrapolation approach is one where there is at best a rudimentary attempt at analytical capacity planning. But after the first or second iteration of the warehouse is created and after the first few users have become captivated of the data warehouse, then careful track is kept for the warehouse and its usage. Over calibrated periods of time, the growth of the warehouse is tracked. Based on the incremental growth that is being measured, an extrapolation of future capacity needs is made. The extrapolation of capacity needs then becomes an educated guess. Of course the educated guess can be refined. The analyst can factor in known growth factors such as the addition of new subject areas, addition of history, and the like. In doing so, the analyst combines the best of the calibrated extrapolation approach and the analytical approach.

But even when the calibrated extrapolation approach is used wisely, the calibrated extrapolation approach has only a short time horizon for effectiveness. Extrapolation can be done for three months or maybe even for six months. But anything beyond that is questionable.

iii. The Third-Party Approach

The Third-Party approach is to find an expert, company or trusted vendor who has worked with a data warehouse, DSS environment that has roughly the same characteristics as your company. There is no substitute for experience. But there are pitfalls with the third-party approach. Some of the pitfalls are:

- the third-party has not provided accurate information;
- the third-party being examined has fundamental business and technological differences from your company;
- the third-party being examined is affected by, and is responding to, business pressures which you are not aware of, etc.

The use of experts and vendors can be beneficial if they have your best interests at heart.

In all cases it must be recognized that the capacity planning effort is an estimate.

References

[1] W. H. Inmon. [Online]. Available: http://www.dmforum.org/portal/library/capacityplanningforthedwInmon.pdf.

[2] W. H. Inmon, "SearchDataManagement," TechTarget, [Online]. Available: https://searchdatamanagement.techtarget.com/news/2240033663/Capacity-planning-for-the-data-warehouse-environment.