Unit 10: Capacity Planning

Calculating Storage Requirement, CPU requirements

Capacity Planning

- Any data warehouse will grow over time, sometimes quite dramatically.
- Every day more data arrives, the total amount of data grows larger, and analysts across the organization are presenting the warehouse with higher volumes of complex queries.
- As the warehouse environment becomes more valuable, capacity planning becomes critical.
- 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.
- 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.

Steps:

- First the tables that will be in the current detailed level of the data warehouse are identified.
- Once the tables are identified, the next calculation is how many rows will there be in each table.
- 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.

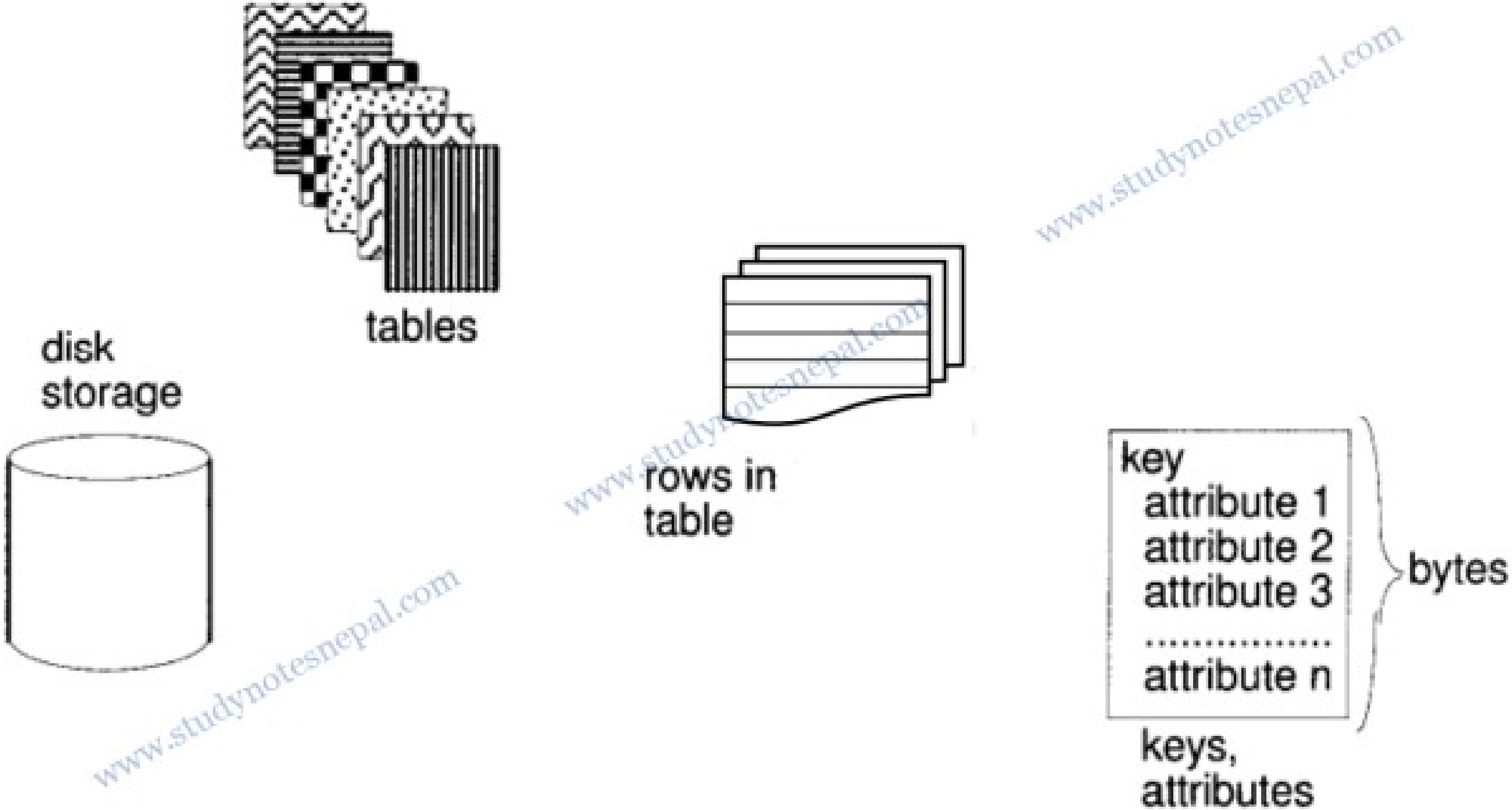

Fig 1: Estimating disk storage requirements for the data warehouse

- Once the contents of the row are taken into consideration, the indexes that are needed are factored in.
- The total disk requirements then are calculated by adding all the requirements mentioned.

Calculating the storage requirements for a data warehouse can be a complex process, but here are the general steps to follow:

- **Determine the total amount of data you need to store:** This involves assessing the volume of data you need to store in your data warehouse. You can do this by estimating the size of each data source or data set you plan to include in the data warehouse.

- **Estimate the growth rate:** You should also consider how much data your organization generates and how fast it grows over time. This will help you plan for future storage needs.

- **Assess compression ratio:** The actual amount of storage needed may vary based on the data compression ratio that can be achieved, as some data types can be compressed more than others. So you need to determine the average compression ratio for the data types you plan to store.

- **Calculate storage space requirements:** To determine the storage requirements, multiply the total amount of data you need to store by the average compression ratio. You may also want to add some additional space for contingencies, such as unexpected growth or data that cannot be compressed.

- **Plan for backups:** Finally, it is important to remember to plan for backups. You should consider the amount of storage space required for backups, how often backups should be performed, and where backups should be stored.

Calculating CPU Requirement

- Calculating the CPU requirement for a data warehouse or data mining project can be a complex process, but here are the general steps to follow:

- **Determine the workload:** The first step is to determine the workload or the types of queries that will be executed on the data warehouse. You should consider the complexity of the queries, the number of concurrent users, and the frequency of query execution.

- **Estimate processing time:** Once you have identified the workload, you should estimate the processing time required for each query. This can be done by executing sample queries or by using benchmark tools.

- **Calculate CPU capacity:** To calculate the CPU capacity required, you need to consider the processing time and the number of concurrent users. You can use the following formula to calculate the CPU capacity:
  - CPU capacity = (processing time * number of concurrent users) / (3600 * utilization)
  - where utilization is the percentage of CPU capacity that is being used at any given time. A typical utilization value is 70% to 80%.

- **Plan for scalability:** You should also plan for scalability and future growth. You can do this by estimating the growth rate of your data warehouse and by using a scalable hardware infrastructure that can accommodate future expansion.

- **Choose appropriate hardware:** Based on your calculations, you can choose appropriate hardware that meets your CPU requirements. This may involve selecting a server with multiple CPUs, high-speed processors, and large amounts of memory.

Approaches for DW Capacity planning:

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

- The Analytical Approach
- The Calibrated Extrapolation Approach
- The Third-Party Approach

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.

1. What are the main parameters to be considered in capacity planning? [BIM 2021, BIM 2018, Group A]
2. What do you mean by capacity planning? Explain how CPU requirement for Datawarehouse is calculated. [BIM 2017, Group B]