TERM PAPER
DATABASE MANAGEMENT SYSTEMS

March 6, 2022

INTERACTIVE DATA EXPLORATION AND VISUALISATION
Roll Number - 19111026
BME/ 6th sem / B.Tech

1   Introduction

A database management system (or DBMS) is essentially nothing more than a computerized data-keeping system. Users of the system are given facilities to perform several kinds of operations on such a system for either manipulation of the data in the database or the management of the database structure itself. Database Management Systems (DBMSs) are categorized according to their data structures or types.

Mainframe sites tend to use a hierarchical model when the data structure (not data values) of the data needed for an application is relatively static. For example, a Bill of Material (BOM) database structure always has a high level assembly part number, and several levels of components with subcomponents. The structure usually has a component forecast, cost, and pricing data, and so on. The structure of the data for a BOM application rarely changes, and new data elements (not values) are rarely identified. An application normally starts at the top with the assembly part number, and goes down to the detail components.

Hierarchical and relational database systems have common benefits. RDBMS has the additional, significant advantage over the hierarchical DB of being non-navigational. By navigational, we mean that in a hierarchical database, the application programmer must know the structure of the database. The program must contain specific logic to navigate from the root segment to the desired child segments containing the desired attributes or elements. The program must still access the intervening segments, even though they are not needed.

2   What is Interactive Data Visualisation ?

Interactive data visualization refers to the use of modern data analysis software that enables users to directly manipulate and explore graphical representations of data. Data

visualization uses visual aids to help analysts efficiently and effectively understand the significance of data. Interactive data visualization software improves upon this concept by incorporating interaction tools that facilitate the modification of the parameters of a data visualization, enabling the user to see more detail, create new insights, generate compelling questions, and capture the full value of the data.

3   2 Challenges and Opportunities

Designing a system for interactive data exploration with a human-in-the-loop front-end requires solving a set of very unique research challenges while also opening the door to several interesting opportunities. In this section, we first outline some of the requirements and challenges, followed by an overview of some of the unique opportunities to address them.

Interactive data exploration has a very unique set of requirements (e.g., response time guarantees), many of which are pushing the boundaries of what is feasible today.

3.1   Interactive Latencies:

By far, the most important challenge in supporting interactive data exploration is to display a result within the latency requirement. Even small delays of more than 500 ms can significantly impact the data exploration process and the number of insights a user makes. Therefore, a new system for IDE need to maintain certain response time guarantees in order to provide a fluid user experience. Moreover, we believe that a system should be able to refine the query answer progressively. This allows users to get a more accurate answer while visually inspecting the query results.

3.2   Conversational Queries:

Different from classical OLAP workloads, users want to explore all different facts of a data set instead of browsing a fixed set of reports. This is very different from what existing analytical databases assume since they expect that the workload is known as prior to create the "right" indexes/samples, whereas the goal of data exploration is to explore and visualize the data in new ways. Moreover, indexes and data cubes suffer from the curse of dimensionality, since memory required is exponential with the number of attributes, making it almost impossible to build an index over all attributes or without knowing the data exploration path ahead of time.

3.3   Rare Data Items:

Data exploration often involves examining the tails of a distribution to view the relatively rare data items. For example, real world datasets are rarely perfectly clean and often contain errors (which are typically rare) that can still have a profound effect on the overall results. Similarly, valid outliers and the tails of the distribution are often of particular interest to users when exploring data (e.g., the few billionaires in a dataset, the super users, the top-k customers, the day with the highest traffic). Unfortunately, for rare events and the tail of the distribution, sampling techniques do not work well since they often miss rare items or require a prior knowledge of the workload, a challenge when designing an system for IDE.

3.4   Connect and Explore:

Ideally, the user should be able to connect to a dataset and immediately start exploring it. However, this requirement implies that there is no time for data preparation and the system has to build all internal storage structures such as indexes on the fly. Another implication of the connect and explore paradigm is that the system has to stream over larger datasets (from the sources) and may not be able to hold the entire dataset in memory (or even on disk). As outlined in the introduction, online aggregation methods are a good fit to overcome this challenge, since they provide an immediate estimate (with error bars) over the incoming stream. However, online aggregation techniques assume that the data is random, which might be false since some data sources (e.g., data warehouses) often sort the data on some attribute. This can result in a biased estimate of the result and invalid error bars. Similarly, no good estimates are possible if the source returns the data in some chronological order and if there is some (unknown) correlation between time and the value of interest (e.g., the sales are increasing over time).