A Literature Review on Column-Oriented Databases

Deepak Kumar Sood
Indraprastha Institute of Information Technology
New Delhi - 110020
Email: deepak15013@iiitd.ac.in

***Abstract*—In today data oriented world, where there is a huge amount of data, retrieving the data fast from the database is a huge problem. Today's row-oriented databases are not efficient for this process. So another type of database is used to handle read-only queries or where the read queries are higher than write queries. Column-oriented database also called as column-stores, reduces the seek time for retrieving the data from the database. The performance of column oriented database is much higher than traditional row-oriented databases. This paper presents an overview of the column-oriented database with its advantages and limitations. This paper ends with an open problem which is presented in future work.**

Keywords: Column-oriented database, row-oriented database, column-store, row-store, vertical and horizontal partitioning.

I. Introduction

Database system performance is directly related to the efficiency of the system at storing data on primary storage (e.g., disk) and moving it into CPU registers for processing. For this reason, there is a long history in the database community of research exploring physical storage alternatives, including sophisticated indexing, materialized views, and vertical and horizontal partitioning. [1] Row-oriented databases are used from the start of the database system implementations because they are easy to use and the changes can be made efficiently and storing the keys and values are easy in this traditional database. But this became the problem when read operations had grown than write operations.

All type of databases may it be row-oriented or column-oriented mostly comprise of a two-dimensional data structure which should be stored in our one-dimensional storage systems as limited by the hardware. Therefore there are two ways to store the data in the linear memory array, one is to store each tuple together as in row-store (group one from each type together) or to store each type of data together as in column-store. Add advantage that the latter provide is that it is faster for analytical processing and is far more efficient in storage space requirement, as same data are stored together they can be compressed using lossless compression by eliminating statistical redundancy.

The figure shows a traditional row-store storage technique vs column-store storage technique, we can clearly see that row-store takes each item of every column and stores it together while on the other hand column-store stores each column data together.

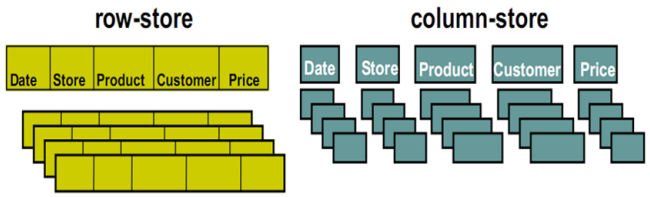

Fig. 1. row-store vs column-store

Section 2 provides all the work that has been done in column-oriented database along with its practical use, limitations and the problems that the technology overcame.

Section 3 describes some open problems that are still in column-oriented database and needs future work.

Finally Section 4 gives the conclusion to this paper.

II. Literature Review

This section will present with the brief overview of the technologies that resulted in today's column-oriented database design along with the limitations it faced, which gave rise to next technology.

TAXIR, implemeted in 1969, was an information retrieval system designed for taxonomic data, to store information accumulated by Plant Introduction about accessions of a major crop, *Phaseolus vulgaris*. [2] This system uses column-oriented database as underlying storage system which was suitable for its data retrieval needs. This database stored the records of the seed crops that come to United States from various places. Each lot had been assigned a unique number and all the data must be handled. That was the time of punched cards and to decrease the resources used and number of operations the data was stored in a column-based structure rather than row-based structure, that was the introduction of column-oriented database system.

The second major implementation that used column-oriented databases was in 1979 by RAPID system developed by Canada Census of population and housing to give statistics on population. [3] RAPID was shared with other statistical organizations throughout the world and used widely in the 1980s. It continued to be used by Statistics Canada until the 1990s. RAPID system contains all the statistical data of the general population in Canada. This data was generally used by

statisticians and economists and to analyze this data the data must be retrieved often. So in this data analytical processing is more important than transactions. But in traditional row-oriented databases analytical processing was slow and gives a huge delay with that kind of large datasets. Therefore again column-oriented database are used than row-based approach.

The next paper titled, "one Size Fits All, Database Architectures Do Not Work For DSS" published in 1995 [4] breaks the news that Decision Support System (DSS) follows "Opposing Laws of Database Physics". At that era traditional row-oriented database ruled with all systems using the same structure for all the work. But in this paper it was proposed that the approach "One size fits all" approach was wrong and for DSS and OLTP operations row-oriented databases do not perform very well.

After this the major paper that presented column oriented database was by Mike Stonebraker et.al. on a paper titled, "C-Store: A Column-Oriented Database". [5] In this paper author proposes a new system C-Store which was substantially faster than any commercial database available at that time. This new system uses column-oriented databases instead of row-oriented databases and use bitmap indexes to complement B-tree structure which increased its transaction capabilities along with analytical processing capabilities. First time a full-scale comparison study was done with other products to get the conclusion that column-oriented database was faster than row-oriented databases.

The implementation and detailed performance analysis is provided in Vibhu Shukla and Dr. Rajdev Tiwari paper, published in IJSR, 2013 named, Column Oriented Database: Implementation and Performance Analysis. [6]

The major differences in row and column oriented database is provided in Abadi Danien J. et.al paper, named, Column-stores vs. row-stores: how different are they really? [7]

III. Future Work

Column-oriented databases are highly efficient for retrieving of data from the database but the main limitation of column-oriented database is its high latency to do a write operation. In banking, column-oriented database are highly used for On-Line Analytical Processing (OLAP) operations but fails in On-Line Transaction Processing (OLTP) operations. So future work should be focused for decreasing the latency for OLTP operations.

IV. Conclusion

The volume of the data is ever increasing in this world as everyday many institutions are moving to digital world, but this also increases the complexity of the database systems and there efficiency decreases because of the limited resources in a system. When the database complexity increases the time needed to retrieve the data from the database increases which can highly effect the user experience and also some other critical functions like in banking and trading sectors where millions of operation are performed in seconds. Therefore column-oriented databases can provide a solution to many operations and can used in many applications in daily life.

References

[1] D. Abadi, P. Boncz, S. Harizopoulos, S. Idreos and S. Madden. The Design and Implementation of Modern Column-Oriented Database Systems. Foundations and Trends in Databases, vol 5, no. 3, pp. 197-280, 2012.

[2] Hudson, L. W., et al. "TAXIR-A biologically oriented information retrieval system as an aid to plant introduction." Economic Botany 25.4 (1971): 401-406.

[3] M. J. Turner, R. Hammond, and P. Cotton. 1979. A DBMS for large statistical databases. In Proceedings of the fifth international conference on Very Large Data Bases - Volume 5 (VLDB '79), Antonio L. Furtado and Howard L. Morgan (Eds.), Vol. 5. VLDB Endowment 319-327.

[4] Clark D. French. 1995. One size fits all database architectures do not work for DSS. In Proceedings of the 1995 ACM SIGMOD international conference on Management of data (SIGMOD '95), Michael Carey and Donovan Schneider (Eds.). ACM, New York, NY, USA, 449-450.

[5] Stonebraker, Mike, et al. "C-store: a column-oriented DBMS." Proceedings of the 31st international conference on Very large data bases. VLDB Endowment, 2005.

[6] Shukla, Vibha, and Rajdev Tiwari. "Column Oriented Database: Implementation and Performance Analysis."

[7] Abadi, Daniel J., Samuel R. Madden, and Nabil Hachem. "Column-stores vs. row-stores: how different are they really?" Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008.