Hinal Bhavesh Desai (ML51414)

09/17/2022

Abstract

The following document gives a description of what are the pros and cons of a row-based vs a columnar-based storage.

Row vs columnar based storage

Research Paper

**Introduction:**

A **data store** is basically a place for storing collections of data, such as a database, a file system, or a directory. In a Database system, they can be stored in two ways. These are as follows: (GeeksforGeeks, 2019)

1. Row Oriented Data Stores
2. Column-Oriented Data Stores

**Row Oriented Databases:**

**Row-oriented databases** are databases that organize data by records, keeping all of the data associated with a record next to each other in memory. Row-oriented databases are the traditional way of organizing data and still provide some key benefits for storing data quickly. They are optimized for reading and writing rows efficiently. In a row store, the data is stored row by row, such that the first column of a row will be next to the last column of the previous row. This data would be stored on a disk in a row-oriented database in order row by row. This allows the database write a row quickly because all that needs to be done to write to it is to tack on another row to the end of the data. Row-oriented databases are still commonly used for Online Transactional Processing (OLTP) style applications since they can manage writes to the database well. Row-oriented databases are fast at retrieving a row or a set of rows but when performing an aggregation it brings extra data (columns) into memory which is slower than only selecting the columns that you are performing the aggregation on. In addition, the number of disks the row-oriented database might need to access is usually larger. Examples of row-oriented databases: Postgres, MySQL. (Blake Barnhill, 2021)

**Columnar Databases:**

**Column-oriented databases** are databases that organize data by field, keeping all of the data associated with a field next to each other in memory. Columnar databases provide performance advantages to querying data. They are optimized for reading and computing on columns efficiently. In a columnar database, the data is stored such that each row of a column will be next to other rows from that same column. These have significant benefits when stored on separate disks. By organizing data by column the number of disks that will need to be visited will be reduced and the amount of extra data that has to be held in memory is minimized. This greatly increases the overall speed of the computation. Examples of column-oriented databases: RedShift, BigQuery, Snowflake. (Blake Barnhill, 2021)

**Row v/s Columnar Databases:**

|  |  |
| --- | --- |
| **Row-Oriented** | **Columnar** |
| Data is stored and retrieved one row at a time and hence could read unnecessary data if some of the data in a row are required. | Data is stored and retrieved in columns and hence it can only able to read only the relevant data if required. |
| Easy to read and write. | Read and write operations are slower. |
| Best suited for online transaction system. | Best suited for online analytical processing. |
| These are not efficient in performing operations applicable to the entire datasets and hence aggregation in row-oriented is an expensive job or operations. | These are efficient in performing operations applicable to the entire dataset and hence enables aggregation over many rows and columns. |
| Typical compression mechanisms which provide less efficient result than what we achieve from column-oriented data stores. | These type of data stores basically permits high compression rates due to little distinct or unique values in columns. |
| Partitioned horizontally. | Partitioned vertically. (Griffith, 2020) |

(GeeksforGeeks, 2019)

**Columnar Databases and Analytical Queries:**

Online Analytical Processing (OLAP) databases facilitate business-intelligence queries. OLAP is a database technology that has been optimized for querying and reporting, instead of processing transactions. OLAP data is also organized hierarchically and stored in cubes instead of tables. It is a sophisticated technology that uses multidimensional structures to provide rapid access to data for analysis. OLAP databases are designed to speed up the retrieval of data. (OLAP, n.d.) Hence an analytical query will struggle when it has to parse a very large number of “blocks” in order to access all of the relevant data. An analytical query will perform an aggregate function on an entire column while ignoring most details about each record and hence analytical query will be far more efficient when data “blocks” are organized by column. (Kaminsky, 2020)

# References

Blake Barnhill, M. D. (2021, August 09). *Row vs Column Oriented Databases*. Retrieved from DataSchool: https://dataschool.com/data-modeling-101/row-vs-column-oriented-databases/

GeeksforGeeks. (2019, August 30). *Difference between Row oriented and Column oriented data stores in DBMS*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/difference-between-row-oriented-and-column-oriented-data-stores-in-dbms/

Griffith, A. (2020, August 10). *Deciding between Row- and Columnar-Stores | Why We Chose Both*. Retrieved from Medium: https://medium.com/bluecore-engineering/deciding-between-row-and-columnar-stores-why-we-chose-both-3a675dab4087

Kaminsky, M. (2020, September 03). *Databases Demystified Chapter 3 – Row Store vs. Column Store*. Retrieved from Fivetran: https://www.fivetran.com/blog/databases-demystified-row-store-vs-column-store

OLAP. (n.d.). *Overview of Online Analytical Processing (OLAP)*. Retrieved from Microsoft: https://support.microsoft.com/en-us/office/overview-of-online-analytical-processing-olap-15d2cdde-f70b-4277-b009-ed732b75fdd6#:~:text=OLAP%20is%20a%20database%20technology,commonly%20stored%20in%20data%20warehouses.