Row Based Datasets VS Column Based Dataset:

Row Based Datasets:

Row-oriented databases organize data by record and retain all of the data associated with a record in memory adjacent to each other. Row-oriented databases are the conventional method of data organization, and they still offer some important advantages for storing data fast. They've been designed to read and write rows quickly. Examples of row based datasets: Oracle, MySQL, Postgres.

The data is saved row by row in a row store, or row oriented database, with the start column of each row next to the last column of the preceding row. This enables the database to quickly write a row because all that is required to do so is append another row to the end of the data. Row Oriented Database records are simple to read and write. Furthermore, Online transaction systems are best served by row-oriented data stores. Typical compression processes produce results that are less efficient than those obtained from column-oriented data storage.

Column Based Dataset:

Column-oriented databases organize data by field and retain all of the data associated with a field in memory adjacent to each other. The popularity of columnar databases has expanded, and they offer performance benefits when querying data. They've been designed to make reading and computing on columns as simple as possible. Examples of column based datasets: Redshift, BigQuery, Snowflake.

Data warehouses were built to help in data analysis. The data is kept in a C-Store, columnar, or column-oriented database so that each row in a column is next to other rows from the same column. Traditional DBMSs retrieve the full row, but column-oriented DBMSs only retrieve the columns defined in the query. Read and write operations are slower in this sort of data storage than in row-oriented data stores. Online analytical processing is better served by column-oriented

storage. These are efficient at performing operations that apply to the entire dataset, allowing for aggregation across a large number of rows and columns (*Compared Analysis of Row-Based Storage and Column-Based Storage*, 2018).

Pros and Cons of Both the Datasets:

Row Oriented Datasets:	**Column Oriented Datasets**
PRO: Data entry and deletion are quick and simple.	**CON:** Inserting and deleting files may have a negative influence on performance.
PRO: Best for transactional processing applications	**PRO:** Best solution for analytical processing.
CON: Data aggregation is a time-consuming and inefficient process.	**PRO:** Best solution for scaling data.
CON: Inadequate compression	**CON:** Incremental data loading is not possible with a columnar database.
CON: More storage is required to store data.	**PRO:** Requires less space to store data

(*Query Based Performance Analysis of Row and Column Storage Data Warehouse*, 2014)

Explain why columnar-based storage is better for analytical workloads?

Columnar dataset dramatically reduces total disk I/O requirements and reduces the amount of data you need to load from disk, columnar storage for database tables is an important aspect in optimizing analytic query speed. When compared to row-wise storage, reading the same number of column field values for the same number of records involves a third of the I/O operations. By utilizing data level parallelism via SIMD instructions, columnar databases can perform faster predicate evaluation. SIMD technology allows multiple column values to be processed in a single CPU instruction. As a result, for analytical applications, a columnar database is favored since it provides for quick retrieval of data columns. Because they scale using distributed clusters

of low-cost technology to boost throughput, columnar databases are ideal for data warehousing and big data processing (Wang & Kogan, 2019).

References:

- *Compared Analysis of Row-Based Storage and Column-Based Storage*. (2018, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9045282?casa_token=a8HNIWBkirYAAAAA:u60n5je-mnkn8edyGyPHJvKXviqHKY2uvRD1lorhBLCBho35V0Kn5jjQWQiVRUedXvKlB_3QI9Y
- *Query based performance analysis of row and column storage data warehouse*. (2014, December 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/7036537?casa_token=SsdiT8mLJw0AAAAA:l9vF3C_lc6ZyT36Z9znCGTZum6VBo12TupF0a6va2Aoe85E9jI3-_bEm_i1CCSdw2Dce0_6wxmg
- Wang, Y., & Kogan, A. (2019). Cloud-Based In-Memory Columnar Database Architecture for Continuous Audit Analytics. *Journal of Information Systems*, *34*(2), 87–107. https://doi.org/10.2308/isys-52531