



Proactive Data Warehousing with Column-Based Databases

Data Warehouses serve as sources of information for analysis, also referred to as online analytical processing (OLAP). Variations in data warehouse design may have a significant performance impact on this analysis. Column-based databases are specifically engineered for OLAP and to deliver high-speed performance. In addition, column-based databases offer lower ownership and maintenance costs than traditional relational databases. Taken together, these factors form the basis for optimum return on investment in Data Warehousing & Business Intelligence.

Javed Syed

PROJECT
performance
CORPORATION

 Part of the AEA group

 please be green - don't print this white-paper unless absolutely necessary





Traditional Databases: The Old Process

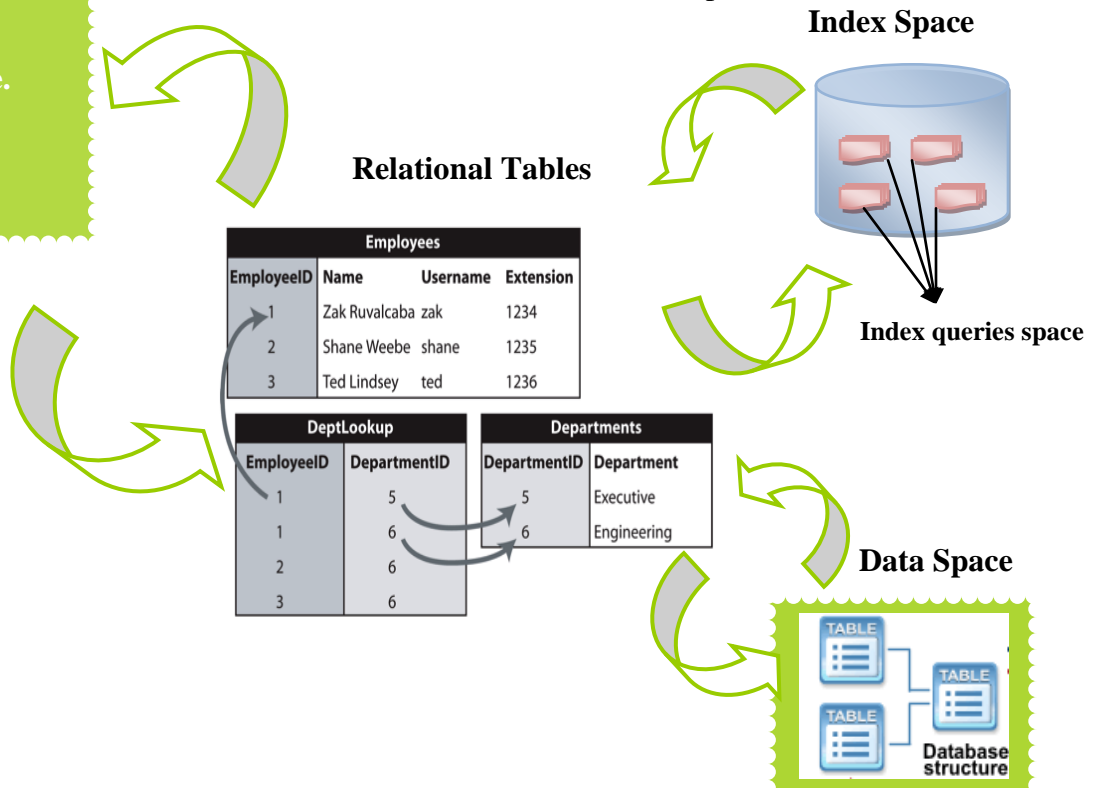
Online transactional processing (OLTP) databases are typically

Disadvantages of RDBMS:

- Designed for only transaction processing.
- Data stored as rows.
- Complex space – consuming indexes.
- Complex to implement on DSS environments.
- High labor-intensive steps.
- More time for loading and data refresh.
- High maintenance costs.
- Tuned to maintain query performance.
- Time-consuming update queries.

used to process day-to-day operations. However, OLTP databases do not perform well for analysis and business intelligence. To perform query-intensive workloads efficiently, OLTP databases require complex, space-consuming index and summary tables. These indexes and summary tables explode data sizes, often requiring five to ten times more data on the reporting system than the original operational system. Traditional relational database management systems (RDBMS) are also more complex to implement for decision support system environments. They require more time to load and refresh reporting systems due to the labor-intensive steps of creating backups, tables, and indexes. There is also a large amount of tuning required to maintain query performance with traditional RDBMS — diagnosing, testing, and tuning queries over and over again.

Traditional relational databases architecture is made up of data space and index space. This architecture works very well for OLTP – POS (Point of Sale) systems, Health Care Claims Processing systems, etc. where records or rows are processed one-by-one. This horizontal approach breaks down for ad hoc queries. Traditional relational databases are not designed to retrieve multiple rows of data or to do aggregation on the fly. This forces database administrators (DBA's) to pre-aggregate data, create materialized views, and create an increased number of indexes (b-tree or bit-map).



Column-based databases compress up to 80% of one petabyte of raw data. Less storage results in fewer servers running, less power used, and overall reduction of CO2 emissions up to 85% over the data warehousing life cycle! This sets a world record for data compression and positions column-based databases as the only “green” databases.

Column-Based Databases: The “Green” Process

Advantages of Column-Based Databases:

- Designed for online analytical processing (OLAP).
- Stores data in columns.
- Index defined on each column.
- Supports vertical partitioning.
- Supports aggressive compression.
- Grid-based parallel database architecture.
- Up to 1000 times faster than RDBMS.
- Query fetches columns that are needed.
- Multiple sort orders.
- Less labor intensive.
- Less time for loading and data refresh.
- Low maintenance cost.
- Update queries are 85% faster.

While standard row-based databases were originally designed for Online Transactional Processing (OLTP), column-based databases have been designed specifically for Online Analytical Processing (OLAP) and ad hoc query analysis. Column-based databases are optimized to deliver higher-speed performance, fewer labor hours, lower cost of ownership, and lower maintenance costs over traditional RDBMS. Column-based databases are highly optimized analytics engines used by many companies for business intelligence, advanced analytics, predictive modeling, regulatory compliance, and rapid reporting.

Column-based databases are designed for the kinds of dynamic business analyses needed to support large numbers of users and large amounts of data. They deliver high-speed access to business information up to 1000 times faster than conventional relational databases.

Traditional relational databases store data by row. Column-based databases store data by column. This is also called Vertical Partitioning. Column-based databases use Bit-Wise™ indexes and do not require data to be pre-aggregated for analysis, allowing users to efficiently and quickly analyze atomic level data.

With column-based databases, queries retrieve only the relevant columns while conventional relational databases retrieve all the bytes in a row, blocking the I/O path and consuming more memory and disk space. Column-based databases fit well with environments where the query loads are time-consuming, and/or more ad hoc queries analysis is required. In this scenario, RDBMS databases cannot be pre-tuned for unexpected queries. But the column-based approach provides effective self-tuning capabilities. In addition, with column-based databases, the business environments which are struggling with performance, space and maintenance issues will benefit not only from faster query results, but they can also conduct more in-depth analysis and reduce the amount of storage by an impressive 70-85 percent.

Column-based databases deliver high-speed access to business information 10, 100, or even 1000 times faster than conventional relational databases.

Column-based database table structure

Healthcare company case study

The old process

This international healthcare company has offices in 38 countries and serves 1.5 million customers worldwide. They process 135 million claims records on an annual basis. The healthcare company was grappling with huge data volume and was struggling to run complex transactions, sub-transactions queries, and report generation.

The existing reporting process was very time consuming and took a whole day to generate reports for 1.5 million claims and customer transaction information. In addition, this existing process was running on an operational system which was built for transaction processing, not for online analytical processing.

The healthcare company received their data from various legacy and operational sources. The data was then pulled through the reporting tool directly from the operational systems

RDBMS Tables Structure

Employees			
EmployeeID	Name	Username	Extension
1	Zak Ruvalcaba	zak	1234
2	Shane Weebe	shane	1235
3	Ted Lindsey	ted	1236

DeptLookup		Departments	
EmployeeID	DepartmentID	DepartmentID	Department
1	5	5	Executive
1	6	6	Engineering
2	6		
3	6		

Column Based Database Tables Structures

EmployeeID	Name	Username	Extension
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

- Indexes stored in columns.
- Queries fetch only needed columns.

EmployeeID	DepartmentID
⋮	⋮
⋮	⋮
⋮	⋮
⋮	⋮

to generate various reports. This resulted in slower performance and often missed deadlines, thus violating the service level agreements (SLA's) for reporting. Moreover, it was very expensive to maintain the extensive support staff needed to continuously work on reporting and on monitoring the operational system.

Analysis of the old process

Our detailed analysis of the healthcare company's old process revealed that the sheer volume of queries driven by demand for reports and for faster, better decision-making had a dramatically negative effect on systems performance.

When companies run reports from the same operational systems that drive the business, slow response times have a direct impact on revenues, productivity, and customer satisfaction. Reports simply can't finish in the time allotted given the resources available. This not only compromises SLA's between IT and business units, but can result in penalties for non-compliance with regulatory agencies. Traditional enterprise

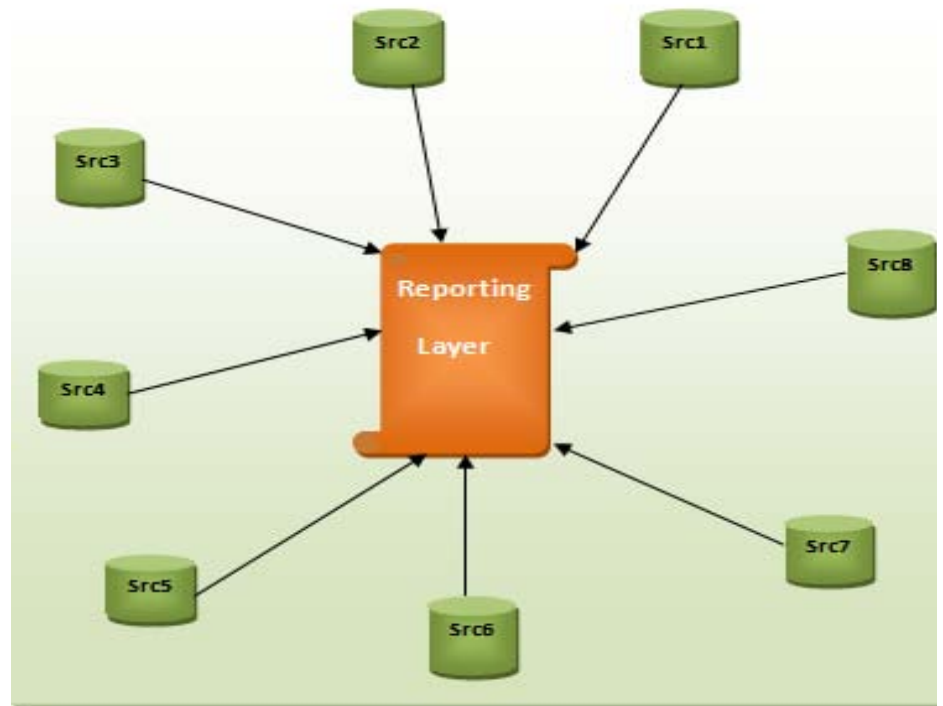


Figure 1.1, shows the old process of the healthcare company.

The healthcare company data warehousing infrastructures with row-based databases are under pressure due to higher ownership and maintenance costs, slower performance, higher space, missed SLA's and reporting windows.

data warehouses or OLTP systems consume large numbers of CPU cycles to read every byte of every row of a large database and deliver the query result. In order to keep performance at target levels, more hardware must be added to the system. Reporting and decision-support also take more DBA time to tune queries, adding indexes and summary tables to ensure acceptable response times.

The business need

To properly address these issues, the healthcare company needed a data warehouse to allow them to rapidly perform ad hoc inquiries. This solution needed to be created within the existing architecture and to provide the ability to easily scale to meet the future growth and current portal requirements. They needed a system to enable simpler and faster queries and sort through

large amounts of data. The health care company was also concerned about the storage and maintenance requirements and the associated ownership costs. As a traditional database expands, the amount of required storage and ownership costs goes up by five to ten times, especially because indexing requires even more space than the database itself. Therefore, this solution was not an option.

Solution

The solution to the healthcare company's problem was to provide the company with a comprehensive data warehousing solution with a cost-effective development plan, minimal startup time, and sound technology. PPC implemented a column-based database to resolve the issues faced by the healthcare company. For choosing the best database to fit business needs, the column-based database emerged as the leader, particularly keeping in mind the ad-hoc queries, space issues and performance statistics.

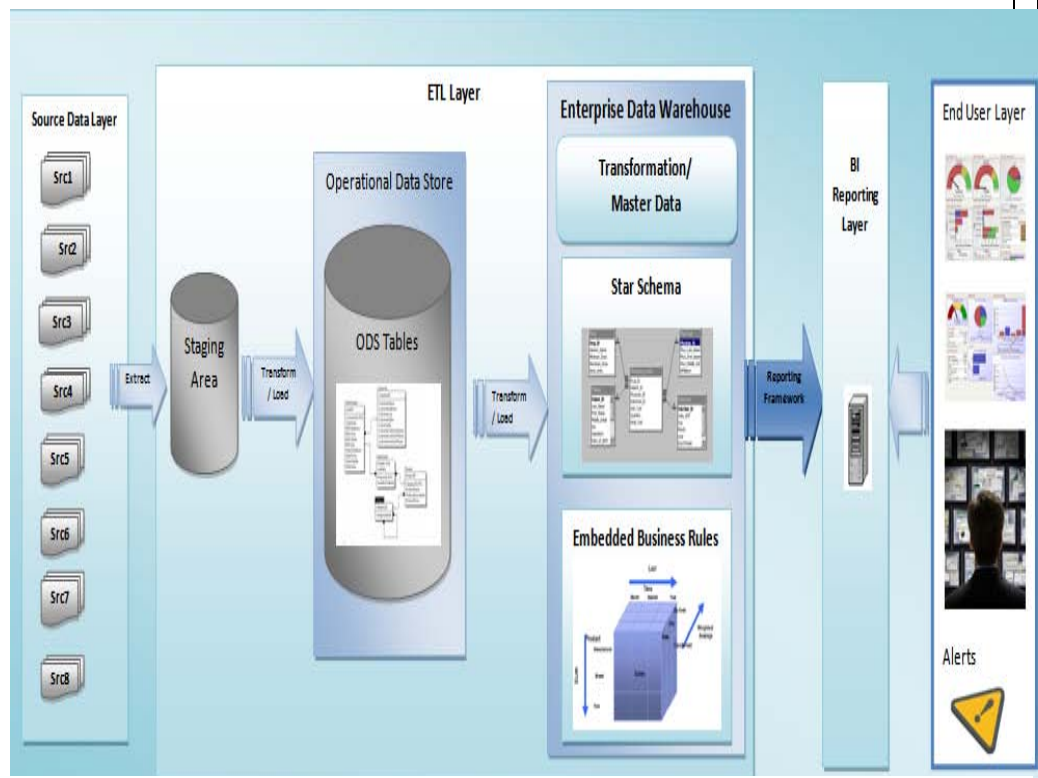


Figure 1.2, shows the new process of the healthcare company with the column based database.

With column based database the healthcare company benefits not only from faster query responses, they can also conduct more in-depth analysis and reduce the amount of storage by an impressive 70%/.

Today, the healthcare company benefits not only from faster query responses, but they can also conduct more in-depth analyses. The previous system could hold only a few months of data, but the new data warehouse stores up to 13 months of data without facing any space constraints. This also provides historical reports over longer time periods.

In addition, legacy systems have been left intact, as a data capture layer has been introduced to retrieve the data from as many as eight source systems. The main function of this process is to collect, consolidate, convert and populate the data into the operational data store (ODS) and then the extraction transformation layer (ETL) is used to feed the data warehouse.



The Right Move: Column-Based Databases

Conclusion:

Organizations need advanced analytics, predictive modeling, rapid reporting, OLAP, and low ownership/ maintenance costs solution to achieve their business goals. Traditional row-based databases fail to address organizations' needs due to slower performance, and higher maintenance/ownership costs. On the other hand, column-based databases are inherently designed for OLAP to provide higher performance with lower ownership and maintenance costs. Furthermore, the unique aggressive compression feature positions column-based databases as environmentally friendly. Therefore, column-based databases are gaining momentum by offering superior solutions to organizations' business and environmental needs.

Project Performance Corporation, part of the AEA group, has experience in implementing column-based databases for private and government clients as part of its Business Intelligence and Data Warehousing solutions. I have described the success story of one of our column-based database project implementations. I hope this paper has helped to clarify the value of column-based database to organizations. For further information on our services and success stories, please visit www.ppc.com.

Corporate Overview

PPC was formed in 1991 with the objective to deliver meaningful solutions to our customers' dynamic problems. PPC has spent last two decades in developing the technology vanguard and our core best practices by which the company drives a higher return on investment for our clients. In November 2007, PPC was named by [Washingtonian](#) magazine as one of the region's great places to work, and it was one of only 10 consulting and contracting firms to be recognized with this honor. In August 2008, PPC joined the forces with the AEA group (A world leading **energy and climate change consultancy**). The following diagram represents well the multiple qualities that have come to characterize PPC's consulting work:



About the Author

Javed Syed is a senior Analyst in the Business Intelligence and Data Warehousing practice at Project Performance Corporation (PPC). Javed has more than 7 yrs of experience in developing and deploying data warehousing, and Business Intelligence solutions. He has experience in Banking, Entertainment, Retail, Government, and Healthcare domains. Prior to joining PPC Javed was involved in building and running corporate data warehouse initiatives, both via onsite-offshore model and as a consultant with Barclays bank, Absa Bank, AOL and TCS. He can be contacted by e-mail, jsyed@ppc.com.

Copyright and disclaimer