Next Generation Data Warehouse and In-Memory Analytics

S. Santhosh Baboo,PhD
Reader
P.G. and Research
Dept. of Computer Science
D.G.Vaishnav College
Chennai – 600106

P Renjith Kumar
Research scholar
Computer Science
Manonmaniam Sundaranar University,
Tirunelveli - 627012
SAP Business Intelligence Consultant,
SAP Labs India

ABSTRACT

Business intelligence is becoming one of the key implementation practices in many industries, with the rapid change in customer buying habits and frequently changing markets every business needs more insight into all the available data that is in the database. Successful business needs to take analytical decision from all possible data that is available. Traditionally the business warehouse system used to get historical data from multiple sources and assist in the decision making process which actually takes time. But to adapt to the rapid changing market there is a need to analyze the business as it happens and take business decisions on real time. With the huge advancement in the hardware and data storage technologies and availability of 64 bit processors, it can help the business intelligence applications to fully utilize the potential of the latest hardware technologies available. The usage of in-memory computing and data storage options like columnar database capability for business intelligence applications can be highly considered for designing the next generation data warehouse systems. This article will analyze the effectiveness of using the in-memory technology for business intelligence based applications and see how it can help in increasing the performance of the business intelligence applications.

Keywords

In-Memory analytics, columnar data storage, Next generation data warehouse, Business intelligence, Real time data analysis, Row store

1. INTRODUCTION

Data warehousing systems involve extraction of huge volume of historical data from various On Line Transaction (OLTP) systems across various complex system landscapes. The traditional Data Warehouse systems extract data from various OLTP source system using Extraction Transformation and Loading (ETL) tools at specified time interval like per day delta load or per hour load basis. But the increasing business complexity demands each business to get real time analytical data based on their existing OLTP system to survive in the frequently changing market and customer buying habits. If business intelligence system can generate knowledge based on the current real time business data rather than reporting on historical data, it will help business to take better and quick business decisions which helps to adopt better with the frequent changing business.

The question is

"Is it possible to design a Data Warehouse system that enables real time computing by bringing Online Transaction processing applications and Online analytical applications at affordable cost using various technologies like In-Memory technology, which give access to real time business process to access instantaneous access to OLTP data?"

If an answer can be for this it will enable business and management users to react to business events more quickly through real –time analysis and reporting based on real time operational data and helps management to support the deployment of innovative new business decisions at appropriate time.

While designing the next generation Data warehouse system there is a need to consider the fact that in current OLTP systems there is huge amount of transaction data including big data which is not used extensively, when an organization can make the analytical business decisions as the business happens in real time, it will be a real swift in the data warehouse and the business decision can help the organization to run even better. To achieve this a business intelligence system need to utilize all the options that take the current data warehouse system to next level like In memory computing, columnar storage and other technologies. By combining the current advances in the hardware technology and having effective in-memory analytics applications there comes a possibility to empower any business by giving the access the real time data and effectively use the real time data for real time analytics there by preventing information lag. These in-memory applications will reduce the total cost of ownership (TCO) and enable faster deployment.

2. BACKGROUND AND DEFINITION

The most popular definition for data warehouse is proposed by Bill Inmon as 'subject-oriented', integrated, time-varying, non-volatile collections of data that is used primarily in organizational decision making".

Ralph Kimball other guru of Data Warehousing defines a Data Warehouse as [1] "A copy of transaction data, specially restructured for queries and analyses"

Hence Business Intelligence (BI) is a broad category of applications and technologies for gathering, storing, analyzing and providing access to data to help enterprise users to make better business decisions. BI Applications includes the activities of decision support systems (DSS), query and

reporting, online analytical processing (OLAP), Statistical analysis, forecasting, and data mining. [2]

In summary Business Intelligence software is a collection of applications needed to make sense of business data. The Data warehouse, a component of this business intelligence tool set is more specific tool responsible for the cleanup, loading and storage of data needed by the business.

A data warehouse can help us to organize the data. It brings together all operative DataSources (Heterogeneous system landscapes) .The job of the Data Warehouse is to provide this data in usable form to the entire organization.

2.1 Data Warehouse architecture

A Data Warehouse consists of several layers. The figure shown below depicts a basic Data Warehouse architecture.

Figure 1 depicts the Data Warehouse layers. There can be many source systems like Enterprise Resource Planning (ERP), Customer Relation Management (CRM) and flat files connected to Data Warehouse system. The Data acquisition layer is connected to the Data Source is system and extracts the raw data into Data Warehouse system. The extracted raw data is kept in corporate memory for further reference. The data is moved further to Data Warehouse layer. The Operational Data store will have the operational data for detailed reporting purpose. This can also be used as other staging layer which can supply the transactional data to other Data Warehouse info providers. Transformation procedures can be applied to transform and cleansing data from one data Warehouse layer to other layer. Once the Operational data is completely transformed to the Info Providers the query created based on these info providers will access the data for reporting purpose.

Figure 1: Data Warehouse layer architecture

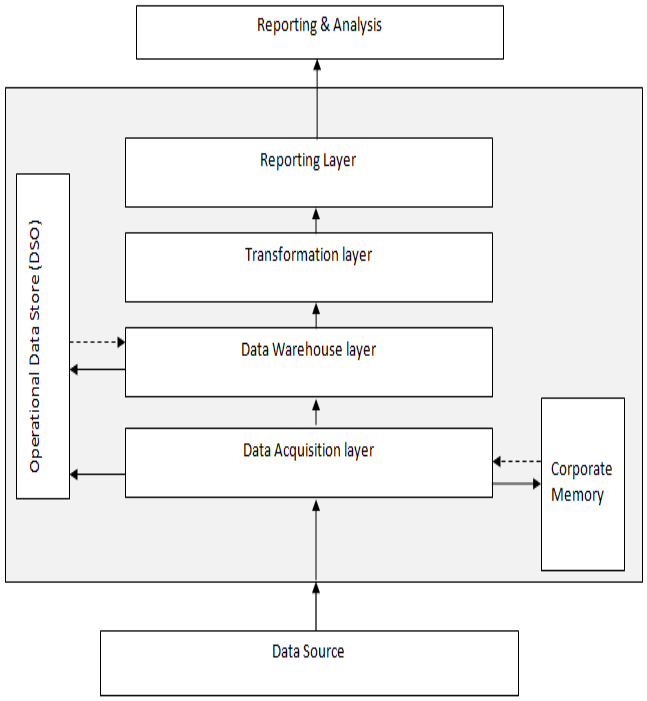
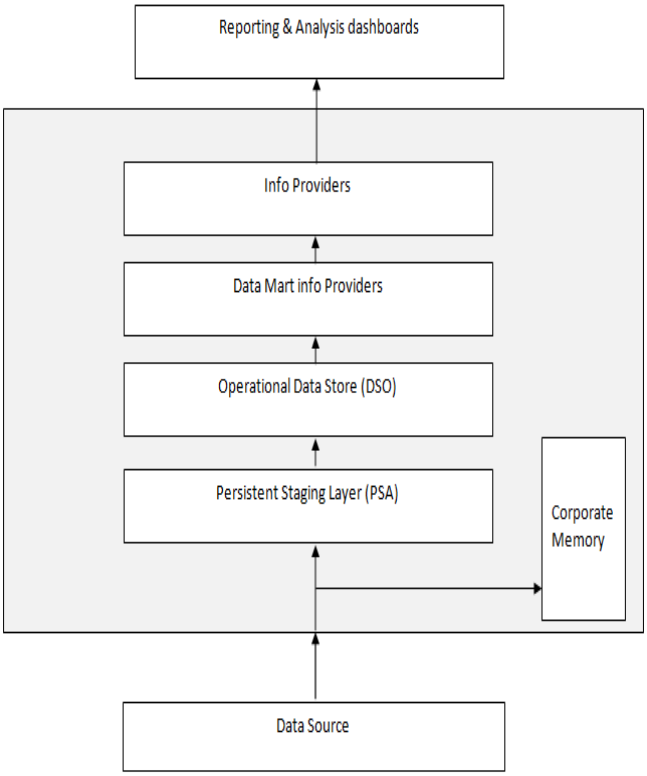


Figure 2: The Data Warehouse objects in each layer



2.2 Performance consideration

In order to have optimal performance and quick response time during analysis operations, data is aggregated during the loads to the reporting layer. Also Business intelligence tools use de-normalized approach like star schema [3] this allows effective read operations on big data volume. In reporting layer the data is often allocated in data marts to serve specific application domain [4]. When there is a need to access only a subset of data from an OLAP cube, the option is to build aggregates based on the selection. It has been claimed that for complex queries OLAP cubes can produce an answer in around 0.1% of the time required for the same query on OLTP relational data. When there is an OLAP cache layer, it acts like a buffer which stores the previous query details. Hence one a query is executed the query runtime will first check OLAP cache and if it is not available it will check if there are any aggregates for the query selection. Still if there is no hit then the OLAP cube data is taken for the query.

3. IN-MEMORY DATABASE SYSTEMS

The conventional Data Base management systems like Relational Database management system (RDBMS) use the physical hard disk drives to store the data. When a query is requested the data from physical hard drive is sent to main memory (RAM) for further processing. In contrast the In-Memory Data base management (IMDBS) stores the data permanently in the main memory of the available hardware. The CPU directly accesses the main memory content. The cost of memory is drastically coming down hence many enterprises that emphasis on the performance of the data loading in Data warehouse system can afford this technology.

[5] In-memory computing allows the processing of massive quantities of real-time data in the main memory of the server, providing immediate results from analyses and transactions.

4. DATABASE DESIGN TRANSITION

With the growth of data stores and business application complexity, solution providers have steadily developed novel approaches to optimizing performance. One key development has been the separation of online transactional processing (OLTP) and online analytical processing (OLAP) systems because both these types of relational databases have somewhat divergent design goals. In the context of BI solutions, they can be considered as follows:

OLTP systems instantly record business events as they happen, such as the sale of a piece of inventory. As such, system designs focus on quickly handling large numbers of small, simultaneous transactions.

OLAP systems provide analysis of the data provided by OLTP systems to support business decisions. They are designed to handle a relatively small number of often complex transactions.

4.1 Row based data storage

Many Enterprise Resource planning applications (ERP) also called as Online Transaction processing (OLTP) Applications have huge volume of transaction data like sales order, purchase order inserted in real time each day. It has day to day transaction data. The database for these kinds of applications prefers write optimized structures.

In Online analytical processing (OLAP) system the aim is to provide insight into the transaction data and assist in making business decisions, the database for these kinds of applications like data warehousing on Business intelligence applications require read optimized data which will help in acquiring better query performance.In row storage, the data sequence consists of the data fields in one table row as shown below.In column storage, the data sequence consists of the entries in one table column.

Table: Country_Sales

	Country	Product	Region	Sales
Row 1	India	Smart phones	South	5000
Row 2	India	Smart phones	North	6000
Row 3	US	Tablets	East	7000
Row 4	Germany	Tablets	West	8000

Row Store view

In this type of storage the contents of each row are stored next in the Database level.

Row 1	India
	Smart Phones
	South
	5000
Row 2	India
	Smart phones
	North
	6000
Row 3	US
	Tablets
	East
	7000
Row 4	Germany
	Tablets
	West
	8000

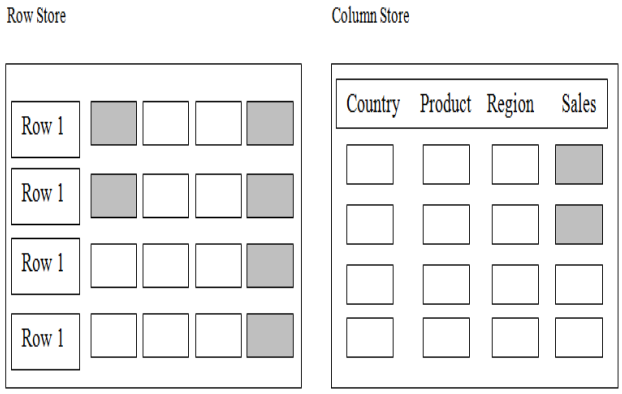
Column Store view

Country	India
	India
	US
	Germany
Product	Smart phones
	Smart phones
	Tablet
	Tablet
Region	South
	North
	East
	West
Sales	5000
	6000
	7000
	8000

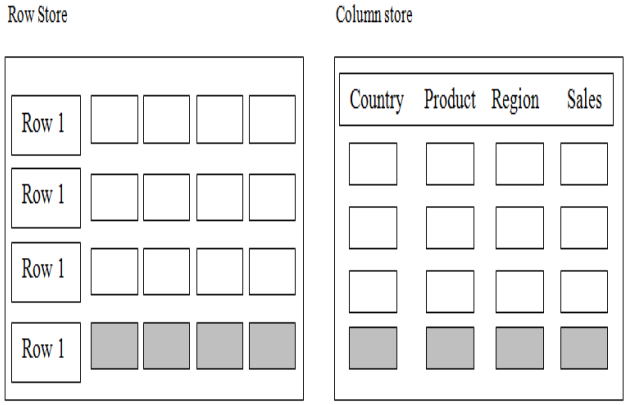
In this storage format the content of the each column is stored in a data sequence

Using SELECT Query on both types of table

SELECT SUM(sales) from country_sales WHERE country = India.



SELECT * FROM country_sales WHERE country = Germany.



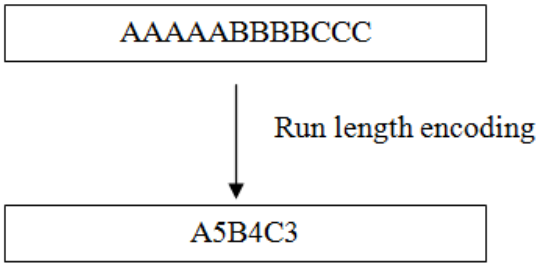
4.2 Compression schemes available

Prefix encoding

Many times when the database have column with predominant value and remaining values in low redundancy, in this type the same value is stored in uncompressed format. In order to handle this, first sort the data set with predominant value, then the attribute vector needs to be started with this predominant value.

Run length encoding

When the attribute vector contains distinct values with many numbers of occurrences, the procedure is to replace the value sequence with single instance of value along with number of occurrences or using offsets based on its starting position.



Usage of columnar layout In-Memory analytics helps the business intelligence systems to use efficient compression techniques. By following techniques like prefix encoding, run length encoding, cluster encoding will benefit the columnar layout very effectively. Also the columnar table layout will enable to do a very fast column scan which helps in avoiding the aggregates and other Data base design complexities.

Results

In a row based storage when there is a query based on OLAP many times it take lot of time because it will hit each and every row and it will make a delay in the overall query performance. The above results show that columnar storage is useful in case of OLAP Online analytical queries as these queries normally get very few attributes from each data entry.

5. IN-MEMORY 64 BIT

There are various factors that contributed to the shift towards In-Memory analytics and products, some of them include high performance hardware resources like multi core processors, parallel servers and increased memory handling capability by hardware. The important fact that will be contributing more in In-Memory analytics is availability of 64 bit operating system due to the 64 bit processors availability with decline In-Memory prices.

[6] Sixty-four-bit systems typically offer faster CPUs and more power-efficiency hardware than older systems. But, for DW professionals, the most compelling benefit of 64-bit systems is the large space of addressable memory.

In-memory databases provide such speed because they don't have disk input/output (IO) to slow them down. The in-memory database is usually a function of a DBMS, but some BI platforms for reporting and analysis also support in-memory data stores and related processing. In many ways, MPP architectures are an alternative to 64-bit-based, in-memory processing, because MPP pools memory resources from many servers (whether 32- or 64-bit) to create a large virtual space.

[5] The theoretical limits of memory capacity are calculated as follows:

32-bit systems: 232 = 4,294,967,296 bytes, or ~4 GB

64-bit systems: 264 = 18,446,744,073,709,551,616 bytes, or ~18 billion GB (18 Exabyte's)

In reality, the upper limits of memory addressability for 64-bit systems are governed by the physical capabilities of the platform and to some extent by the cost of the memory itself. As platforms become more capable, system memory capacity continues to grow, and at the same time, the cost of memory on a per-gigabyte basis continues to drop. Server memory capacities are now often measured in terabytes (TB), where 1 TB = 1,000 GB.

6. PARALLEL PROCESSING WITH IN-MEMORY COMPUTING

When the in-memory analytics application considers the various software layer optimizations for parallelization the speed can be still increased.

Amdahl's law

Gene Amdahl conducted fundamental considerations about software-level parallelism. He defined that the maximum speed up of executing a piece of code in parallel is limited by the time needed to process the longest sequential fraction of the code. This is nowadays known as Amdahl's law [7]

In the case of parallelization, Amdahl's law states that if P is the proportion of a program that can be made parallel (i.e., benefit from parallelization), and (1 − P) is the proportion that cannot be parallelized (remains serial), then the maximum speedup that can be achieved by using N processors is [7]

$$\text{max. speedup}(N) = \frac{1}{(1-P) + \frac{P}{N}}$$

In the limit, as N tends to infinity, the maximum speedup tends to 1 / (1 − P). In practice, performance to price ratio falls rapidly as N is increased once there is even a small component of (1 − P).

Hence the best practice for the effective parallel processing with in-memory computing lies in attempting to reduce the component (1 – P) to the smallest possible value.

As an example, if P is 90%, then (1 − P) is 10%, and the problem can be sped up by a maximum of a factor of 10, no matter how large the value of N used.

7. IN-MEMORY ANALYTICS

[8] In-memory analytics refers to the approach of aggregating, querying and analyzing data when it resides in a computers random access memory (RAM), as opposed to querying data stored on physical disks. In-memory eliminates the traditional disk access bottleneck inherent to traditional databases when performing analytics delivering significantly increased performance.

This new analytics method will be of very helpful to business users, decision makers and analysts to get various insights into the business based on the data, with much improved performance the decision making will be very quick and adapt to the quickly changing business trends. With in-memory in 64 bit by executing things in parallel with advantage of multi core chips business intelligence systems get the best ever processing speed for the analytics applications which helps in better decision making. Also with the In-Memory there is less dependency for storing the data in OLAP cubes or various aggregates, This reduce the implementation cost and total cost of ownership.

Currently the query performance is increased by these measures currently

- Building aggregates on the available cube
- Creating Index on the large tables
- OLAP cache is referred first during the query execution

With In-Memory analytics the Business intelligence systems can provide extremely fast query response times, it also helps to create the index at the data base level. It helps to avoid building aggregate table based on the cubes which consume space. Since the data reside in RAM the query and report generation will require less network access and disk I/O. With all these features an organization can definitely reduce the total cost of ownership and there will be less maintenance efforts with increased query performance which helps to make better business decisions and provide valuable business insights and helps each business to get fine-grained analysis with the huge volume of data available. Also systems can query with extremely fast response time even on big data.

Benefits of in- memory analytics

- Increase in performance which can be millions of times faster than disk based access
- This can also be used by small and medium industries as the set up and maintenance is simpler
- We can take business decisions as the business happens on real time.

Hardware factors that contribute towards more effective In-Memory analytics

- When the core count like 8 core is increased to 10 cores in the processor it provides more computation resource capability with In-Memory analytics
- When there is more memory capacity says 2 TB for four ways server it will help to hold large volume of data which can be utilized by the query.
- Increase of cache up to 30 MB can still help in holding more data in cache than reading from RAM which will have further performance improvement.

Intel XEON processor E7 family has these capabilities which support effective In-Memory analytics.

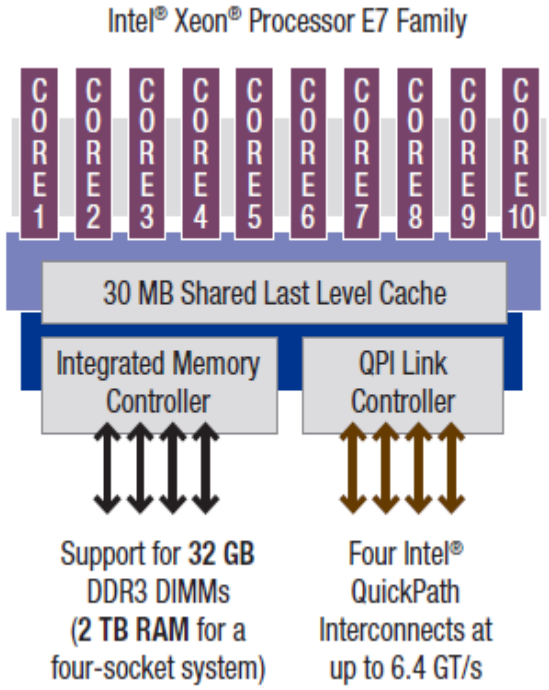


Figure 3: From Analyzing business as it happens, Whitepaper, Intel & SAP, Version 1.0 April 2011 [5]

8. FUTURE WORK

Until now the standard procedure is to use OLTP and OLAP as separate systems to overcome the performance and resource management issues. The data is loaded on a regular interval into OLAP system and business decisions are taken after the data is loaded. Since there is tremendous hardware advancement there seems to be bright option to club OLTP and OLAP in a single system, the option of using in-memory computing methodology will be considered, if this can be achieved there will not be any need for dedicated business intelligence servers that extracts data from OLTP system and since OLTP and OLAP are integrated and business intelligence systems can take the analytical data directly from the underlying OLTP system in real time.

9. CONCLUSION

In this paper a detailed study is done about the data warehouse system architecture in multiple views, the use of in-memory technology and columnar storage is analyzed in detail with respect to analytics applications. With all these study it can be clearly concluded that having capability of providing extremely fast query response time, in- memory analytics will help organization's to get better insights into the huge volume of business transaction data. When compared to the traditional row based storage, column based storage along with in-memory analytics have more advantage due to the fact that the OLAP queries will fetch data only for few attributes of each data entry. The next generation of Data warehousing system and business intelligence analytics application will be designed definitely based out of in-memory applications which help the organization to adapt to the quicker change in business trends with more emphasis on real time reporting and more and deeper insights into analytical data there by enabling organization to take effective business decisions.

10. REFERENCES

[1] R. Kimball "The Data Warehouse toolkit, 1996, page 310

[2] http://searchdatamanagement.techtarget.com/definition/business-intelligence

[3] Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme - Einordnung und Überblick. In: Analytische Informationssysteme, Springer, Berlin [u.a.], 1998, pp. 3–25.

[4] Surajit Chaudhuri and Umeshwar Dayal (1997). "An overview of data warehousing and OLAP technology".

[5] Whitepaper: Analyzing business as it happens, Intel & SAP April 2011

[6] Next generation Data Warehouse platforms Philip Russom, TDWI

[7] G.M. Amdahl, "Validity of the Single-Processor Approach to Achieving Large-Scale Computing Capabilities," Proc.Am. Federation of Information Processing Societies Conf., AFIPS Press, 1967, pp. 483-485

[8] Seema Haji, Dec 16, 20122 , March to in-memory http://www.birst.com/blog/2012-march-memory