Data warehousing technology trends

Data warehouses are no longer simple decision-support databases fed by batch extract, transform and load (ETL) processes. They’ve evolved into dynamic analytical warehouses, and technological innovations are rapidly changing data warehouse architectures and providing the potential for substantial performance improvements. In this e-book, get an update on the most important trends in data warehousing, including appliance technology, columnar databases, in-database analytics and more.

Sponsored By: HP
Data warehousing technology trends

Table of Contents

Selecting the right data warehouse platform for your organization
Understanding in-database analytics technology: Benefits, uses and ROI
Real-world tips for improving BI performance, data warehousing systems
Resources from Hewlett-Packard Company
Selecting the right data warehouse platform for your organization

By William McKnight, SearchDataManagement.com Contributor

As companies consider their ability to manage the asset that is information, the data warehouse platform (and its database management system, or DBMS) is the most important decision point. The platform is the foundational component of the tool selections, the consultancy hires, the architecture, etc. In short, it defines your information culture.

However, in selecting the platform to support the data warehouse, organizations are faced with an exponentially higher number of variations and distinct departures from the traditional online transactional processing (OLTP) database management systems than ever before.

Over time, data warehouse data volumes will continue to soar as organizational history accumulates, syndicated data is collected and new sources with more detailed data are added. Furthermore, the user community consuming the data continues to grow, expanding well beyond company boundaries to customers, supply chain partners and even users on the Internet. Companies need to make sure they choose a proven platform not just for the initial known requirements but also one with the ability to scale to meet future, to-be-determined requirements.

Data warehouses, to be successful, need to provide:

- Quality data that is available to users when they need it.
- An architecture that enables low long-term total cost of ownership.
- Good query performance that results in increased interactive usage.
- The ability to integrate real-time data feeds.
- A platform to support mixed and unpredictable workloads.
- A scalable path forward as data needs grow.
Criteria for data warehouse platform selection

The decision process for selecting a data warehouse platform should go well beyond the usual consideration of the operational DBMS vendor. Today, that decision can be nuanced along several potential requirements, including:

- Active loading of data and immediate access to loaded data.
- A mixed processing workload against the data.
- Cross-functional complexity.
- The desired level of query concurrency.
- The organization’s platform scalability needs.
- The required DBMS functionality.

Given the state of the marketplace, the technical architecture for a data warehouse platform should be:

**Scalable.** Your platform should be able to scale in both performance capacity and incremental data volume growth. Make sure the proposed platform scales in a near-linear fashion and behaves consistently with growth in database size as well as the number of concurrent users and the complexity of queries. Understand the additional hardware and software required for each of the incremental uses.

**Powerful.** The platform should be designed for complex decision-support activity in a multiuser, mixed-workload environment. Check on the maturity of the query optimizer for supporting all type of queries with good performance and to determine the best execution plan based on changing data demographics. Check on conditional parallelism and the causes of any variations in the parallelism deployed. Check on dynamic and controllable prioritization of resources for queries.

**Manageable.** The technology you choose should need only minimal support tasks requiring database administrator (DBA) or systems administrator intervention. It should provide a single point of control to simplify system administration, and you should be able to create and implement new tables and indexes at will. In addition, it should include a robust set of management features and functions, including DBA productivity tools, monitoring utilities,
locking schemes and other security mechanisms, remote maintenance capabilities and user chargeback functionality.

**Extensible.** The data warehouse platform should provide a flexible database design and system architecture that can keep pace with evolving business requirements and that leverages existing investments in hardware and applications. Ask such questions as: What is required to add and delete columns? What is the impact of repartitioning tables?

**Available.** The platform should support mission-critical business applications with minimal downtime. Check on “hot pluggable” components, and understand your system downtime requirements and any technical issues that might deny or degrade service to end users. That can include batch load times, software/ hardware upgrades, severe system performance issues and system maintenance outages.

**Interoperable.** You should choose a data warehouse platform that allows for integrated access to data on the Web, internal networks and corporate mainframes.

**Affordable.** The proposed technology – hardware and software, plus services and required customer support fees – should provide a low total cost of ownership (TCO) over a multiyear period.

**Flexible.** The platform should be able to provide optimal performance across the full range of normalized, star and hybrid data schemas with large numbers of tables. Look for a proven ability to support multiple applications from different business units, leveraging data that is integrated across business functions and subject areas.

**Referenceable.** There may not be a single customer reference that matches your environment exactly, but you should be able to see a consistent trend across a wide range of references that reinforces what you’re looking for in a data warehouse platform.

Organizations should also consider vendor viability, especially in these days of marketplace consolidation. The financial stability of vendors, the importance of data warehousing to their overall business strategies and their continued spending on research and development in
these areas toward a well-developed and relevant technology vision are all key components of this critical decision.

**Data warehouse architecture considerations**

To enable the above criteria, companies have some new options in the types of data warehouses that are available, many of which have been created in recent years owing to the high TCO and potential limitations of traditional data warehousing approaches.

Typically, data warehouse environments have several restrictions that organizations have "learned to live with" in seeking acceptable analytical performance. Indexes, summary tables, cubes, utility executions (i.e., reorganizations) and various de-normalizations are often introduced into the environment to help boost performance. There's a point at which overcoming these challenges ceases to be "easy" and affordable and therefore ceases to get done. But perhaps more important than these very real restrictions are the restrictions that they can lead to in “possibilities thinking” and information exploitation. The higher the untuned performance of a data warehouse environment is, the better.

The following technologies may help organizations avoid some of the performance restrictions:

**Columnar databases.** A row-oriented database is an implementation of relational theory. Data is stored as bytes, with all of the columns in a row stored in order. These bytes are grouped by the several thousand – from 4,000 to 64,000 – into data blocks. This is the unit of input/output, an exception being the “pre-fetch” capability of a row-oriented DBMS to sense a pattern in the data reads. A columnar DBMS also is an implementation of relational theory, but with a twist: The data storage layer contains a grouping of columns. For example, all of the Column 1 values are physically stored together, followed by all of the Column 2 values, etc. In addition, the data is stored in record order, so the 100th entry for Column 1 and the 100th entry for Column 2 belong to the same input record.

In a columnar database, all of the same data – your data – is there. It’s just organized differently (automatically, by the DBMS). The main benefits include the ability to highly
compress the data. A byproduct of the increased compression is the ability to perform columnar operations – like MIN, MAX, SUM, COUNT and AVG – very rapidly.

**Data warehouse appliances.** Data warehouse appliances have emerged as viable short-list solutions for new deployments or refurbished data warehousing installations. Appliance vendors now dot the landscape (with representation in analyst "quadrants" and "spectrums" plus hundreds of millions of dollars in venture backing), and the top data warehousing vendors have jumped into the appliance business through internal development or acquisitions. The devices are worthy of any data warehouse professional’s understanding, attention and consideration.

The data warehouse appliance is a combination of hardware, software, operating system, DBMS and storage pre-configured for data management requirements and uses. Many utilize commodity components, and some include open source DBMS software. The open source technology provides a starting point for basic database functionality, and appliance vendors focus on necessary functionality enhancements. Query performance, especially against large volumes of data, is distinctively impressive thanks to the automatic parallelism that many, though not all, appliances provide. Low TCO for a mixed-workload data warehouse environment is also possible, and consequential, with appliances.

**Parallel architecture.** Parallelism is as important as columnar orientation to query performance success, and massively parallel processing (MPP) systems provide the most parallelism possible. Sharing disk and/or memory across nodes creates overhead. Sharing nothing minimizes disk access bottlenecks and is preferred in MPP, which also features an exponential uptake from symmetric multiprocessing systems and clusters in the number of processors supported and the sophistication of the interconnect. MPP is the generally acknowledged ideal parallel architecture for analytic querying and other data warehousing and business intelligence (BI) applications.

**Packaged analytics.** Finally, most large enterprise software vendors now offer some type of pre-built reports or dashboards centered on typical business functions such as finance, sales and marketing, and supply chain management. In addition, a number of smaller companies have developed packaged analytics built on licensed platforms and targeted to
specific industries, applications or specialized functions, or to supplement other third-party services.

While a true “data warehouse in a box” has remained somewhat elusive, analytic packages bridge the gap between off-the-shelf canned reports and heavily customized BI software. Such packages simplify the development process and deliver functionality with more cost certainty. Organizations must complete a careful evaluation process to determine whether a package is right for them, and they must understand the various costs and benefits of different options. With all of this information at hand, decision makers will be better positioned to make the best choice for their organizations.

**Data warehouse selection process**

Once the possibilities have been narrowed down according to an evaluation of the available technologies and options, a proof of concept (POC) project should be done with your data. Data warehouse vendors all have the ability to take in your data for an off-site POC. These off-site data centers are well worth using; you can “visit your data” in person during the process, if you wish.

You need to locate and then gather representative data as well as determine the data access paths that you wish to explore in a POC. The proof of concept should comprise these phases:

- Prepare and gather data.
- Set up test environment(s) at off-site location(s).
- Validate data-access tool connectivity, performance and load capabilities, with an emphasis on concurrency.
- Compile qualitative information on each offering being considered, using published information, consulting help and customer references.
- Weigh quantitative and qualitative recommendations on each vendor.
- Make sure each important criterion is covered in the POC.

These days, the best answer to your data warehouse platform selection needs may not be as evident as it was in the past. Given the high relevance of data and information,
investment and innovation continue in this area and there are a surprisingly large number of options. Many platforms may work but at high short- and long-term prices. Making a selection consistent with your goals as an organization and with a broad and detailed view of the landscape is paramount to success.

**About the author:** William McKnight is the president of McKnight Consulting Group. He works as a strategist, lead enterprise information architect and program manager for complex, high-volume and full-lifecycle implementations worldwide within the disciplines of data warehousing, master data management, business intelligence, data quality and operational BI. Implementations by his teams have won IT best-practices awards. McKnight is a Southwest Entrepreneur of the Year award finalist and a frequent best-practices judge; in addition, he has authored more than 150 articles and white papers and given over 150 keynote speeches and public seminars. He holds an MBA and is a former vice president of IT at a Fortune 50 company and a former DB2 engineer at IBM. McKnight is also the author of the book 90 Days to Success in Consulting. He can be reached at william@williammcknight.com.
Understanding in-database analytics technology: Benefits, uses and ROI

By Elisabeth Horwitt, SearchDataManagement.com Contributor

In-database analytics is an emerging practice that experts say can significantly cut the cost and time it takes to do complex and data-intensive analytic processes.

This article will cover how in-database analytics works and how it differs from more traditional analytical methodologies. It will also look at recent and developing market trends, such as vendor support and the emergence of open standards, as well as the potential of in-database analytics for revolutionizing advanced analytics and business intelligence (BI).

Breaking down in-database analytics

According to a November 2009 report from consulting firm Forrester Research Inc., titled "In-Database Analytics: Heart of the Predictive Enterprise," the practice is far from bleeding edge. In fact, in-database analytics is the latest instance of a longstanding approach in which developers embed application logic into data warehouse and database systems.

In a traditional setup, predictive analytics, data mining and other compute-intensive analytic functions are part of separate applications or data marts, each typically with its own system, set of data, analytic tools and programmers. As a result, "a lot of people spend a lot of time shepherding data out of a database, profiling it, transforming into a format a particular analytic tool can digest, and moving it to where analysts can use it," says Neil Raden, president of Santa Barbara, Calif.-based consultancy Hired Brains Inc.

In contrast, with in-database analytics, the analytic functions reside on the same centralized enterprise data warehouse (EDW). This eliminates I/O-intensive extract, transform and load (ETL) operations that can consume as much as 75% of cycle time in predictive analytics applications. It also enables developers to exploit powerful data warehouse platform technologies, such as parallel processing.
Empowering the enterprise data warehouse

In-database analytics is one of several recent developments that have made advanced analytics an increasingly important, and affordable, element of corporate BI initiatives.

First, a precipitous drop in storage, computing and memory prices has helped fuel the emergence of scaled-down, low-cost database engines, data warehousing platforms and appliances.

Second, many of these platforms support leading-edge computing technologies that enable compute-intensive applications like advanced analytics to run more efficiently. For example, 64-bit memory enables large volumes of data needed for predictive analytics to reside in main memory instead of on disk, which eliminates time-consuming I/O transfers. Parallel processing enables multiple analytic processes to run in tandem. Virtualization enables companies to allocate computing resources to analytic and database querying functions on a prioritized and as-needed basis.

Another key factor is the recent emergence of two industry standards for advanced analytics. MapReduce, a vendor-neutral programmability framework for complex information types, is gaining traction among data warehouse and advanced analytics software vendors, according to Forrester.

The second standard, Hadoop, defines an open analytic processing pushdown workflow model and distributed analytic object-file store. It also has growing support from database, data warehouse and cloud computing platform vendors, Forrester said.

Once these standards gain broad support from leading players, businesses will have far more flexibility in choosing (and migrating between) data warehousing platforms.

At least as important, MapReduce and Hadoop can work with unstructured as well as structured data residing in a database. This will be critical for the next generation of analytic applications, which will be mining the complex patterns in diverse and distributed information generated by Web 2.0 applications, social networking, clickstream analysis and
the like, said Forrester analyst James Kobielus, lead author of the report.

**Who can benefit from in-database analytics?**

In-database analytics is potentially useful to many types of organizations pursuing advanced analytics. It's well-suited for activities such as targeted response marketing, dynamic pricing analysis and fraud detection and prevention, according to analysts. It can also help executives who need to know how best to allocate R&D money or funding for security upgrades; or who need to create a business plan that reacts to projected market changes over the next five years.

Still, it isn't for everybody, Raden warned. He recommended that companies considering whether to deploy in-database analytics, either as an in-house development platform or to support commercial applications, should ask themselves the following questions:

- Do you have any business problems that lend themselves to predictive modeling?
- Is the necessary data readily available, consistent, accurate and supported?
- Above all, do you have the corporate culture and the will to make use of the results you obtain?

In other words: "Are you willing to let math algorithms in a computer lead you to do things you didn't conceive of yourself, and go against what you're already doing?" Raden asked. "A lot of companies aren't there yet."
Real-world tips for improving BI performance, data warehousing systems

By Craig Stedman, SearchDataManagement.com Site Editor

Organizations are continually looking for ways to improve the performance of their data warehousing and business intelligence (BI) systems. And data warehousing and BI performance improvements are becoming even more important as companies broaden their deployments of BI tools to include more business users and emerging concepts such as operational BI and real-time BI take hold.

In this interview, Rick Sherman, founder of consulting firm Athena IT Solutions, provides practical advice on how to get faster performance from data warehousing and BI systems, both during the development process and after they’ve been deployed. Sherman also offers guidance on getting executive buy-in and funding for performance improvement initiatives, including purchases of new technologies.

What are some of the common performance bottlenecks that can slow down data warehousing and BI systems?

I’d break this up into two areas. The first would be the data loading – people often have trouble with integrating the data, cleansing the data and getting the data into the warehouse in a timely manner. Usually there’s a small window of opportunity for them to pull data from their source systems, and they struggle with getting that data in there.

On the front end, the frustrations that business users have – first off, they often get frustrated with just starting up reports. They get on the [corporate intranet], they click on a report – sometimes it takes quite a while for that report to get refreshed. Second, business people love dashboards, but they click on a graphic to get more detailed data, and then it takes an inordinate amount of time to get the detail and drill down and really try to discover what’s going on. And a third area that’s very common is with deeper analysis, when people are looking at multiple dimensions – products, geographies, other business metrics – and it takes them a very long time to get the results of those [queries].
Can you give people some advice on what they can do upfront, when they’re building data warehousing and BI systems, to avoid performance problems later on?

The first [thing] is to establish a data and information architecture and set it up for specific purposes. Set up a warehouse for data integration, to load and cleanse the data – tune it for that, and don’t let anybody go and do reporting off of it. The second [part of that] is to set up a distribution database, so that the data can be spread out to different data marts, cubes or other downstream applications. And the third is to set up databases or applications for specific types of analytics. You need to realize that it isn’t one-size-fits-all. This is a common mistake I see, and it’s one of the reasons business users get frustrated. People design the cubes or the databases to answer every question that somebody in the enterprise might want to ask, and that really bogs down any kind of reporting or analytics. What you need to do is separate out the types of analysis that somebody is going to do, and then also split up the data, into various pockets – sort of the divide and conquer strategy. Then the tools work much more effectively.

Now let’s say I have an existing data warehousing and BI system and performance has become an issue. What sort of things can I do to improve it?

This is again a case where you probably want to split it into [different problems]. Is it a data loading and data integration issue? Or is it on the analytics, reporting and query processing side? On the data loading side, the good news for folks is that there are many ETL or data integration tools that can load data very effectively. And there are some different file utilities that are used by ETL tools to help speed up data [loading] in the background. It’s not so much going out and getting new tools on the back end, but better utilizing the tools [you have]. Or if you’re picking tools, look for things like in-memory caching and other techniques for loading data faster.
What about technologies such as in-database analytics, in-memory analytics, data virtualization and data warehouse appliances? Can they boost the performance of BI applications, and are they mature enough for widespread use?

Just collectively, I’ll say that they all will boost performance if used correctly. But let’s get into each one. In-memory analytics is certainly in prime time; most of the large BI vendors and even some of the open source vendors have incorporated various aspects of in-memory analytics into their tools. With 64-bit architectures and the amount of memory that is available on servers now, this is a very effective strategy to improve throughput. When I was at a database company a number of years ago, the trick was always not to write things to disk, to keep things in memory, and that still is true today.

In-database analytics is really geared toward: “OK, we know the usage patterns and how data is going to be analyzed and joined together, so let’s structure the database in order to be able to handle that.” It’s similar to the way that some of the multidimensional OLAP databases were created 10 or 15 years ago to boost performance. This is sort of those OLAP cubes, only on steroids. Data virtualization is another technology that offers some pretty good [performance] advantages, especially with real-time analytics and operational BI. Rather than moving data to a warehouse, in real time it does data integration and provides you with the data. There are some limitations with the amount of data that you can do and how many systems you can touch. But in call center applications or [other] areas where you’re trying to get data about a specific customer or a specific order, it works out very well. It’s a good complement to a data warehouse, a good way to bridge the gap if you don’t have data in your data warehouse.

Another [potential option] certainly is in BI and data warehouse appliances. With software vendors buying hardware vendors, with BI appliance companies being bought out by the high-tech titans, [appliances] are definitely going to get more and more pervasive in the marketplace.
And what are some of the steps that data warehouse and BI managers should take to build a business case for purchases of new technologies or other performance improvement actions?

I think the issue gets to be, how much of a pain is it? If the business users have been struggling to get data and analyze it on a timely basis, if they’ve been building “data shadow systems” or spreadmarts in order to be able to do the analysis, those are prime candidates where it won’t take very much to get the business to sign up to provide an alternative – either other kinds of technologies or other kinds of approaches. That’s almost a lay-down – it’s a pretty easy case [to make].

The second one is tougher. That’s where there are some performance issues, and the business is sort of frustrated but not really willing to make a big investment in [fixing the problems]. You can talk to them about various technologies, but data virtualization or in-memory analytics isn’t the sexy kind of thing that makes a business person say, ‘Yeah, I need that.’ In that particular case, it’s almost a take-it-slow approach, where what you might want to consider is setting up a pilot or a proof of concept with some key business people. Once they get excited about it, you can use them as evangelists to expand it out. But you really need to show the business that these things can work out and that they’re going to get the performance and the analytic capabilities they really want.
Resources from Hewlett-Packard Company

When Your OLTP Applications Need Extreme Performance

Vertica Analytics Platform Release 5.0 Overview

About Hewlett-Packard Company

HP is a technology company that operates in more than 170 countries around the world. We explore how technology and services can help people and companies address their problems and challenges, and realize their possibilities, aspirations and dreams. We apply new thinking and ideas to create more simple, valuable and trusted experiences with technology, continuously improving the way our customers live and work.

No other company offers as complete a technology product portfolio as HP. We provide infrastructure and business offerings that span from handheld devices to some of the world's most powerful supercomputer installations. This comprehensive business portfolio helps us match the right products, services and solutions to our customers' specific needs.