BeyeNETWORK

Global coverage of the business intelligence ecosystem

Analytic Platforms: Beyond the Traditional Data Warehouse

By Merv Adrian and Colin White

BeyeNETWORK Custom Research Report
Prepared for Vertica

Executive Summary

The once staid and settled database market has been disrupted by an upwelling of new entrants targeting use cases that have nothing to do with transaction processing. Focused on making more sophisticated, real-time business analysis available to more simultaneous users on larger, richer sets of data, these analytic database management system (ADBMS) players have sought to upend the notion that one database is sufficient for all storage and usage of corporate information. They have evangelized and successfully introduced the **analytic platform** and proven its value.

A dozen or more new products—the majority introduced after 2005—have been launched to join the pioneering analytics-specific offerings, each of which boasts thousands of installations. Collectively, the newcomers successfully placed an additional thousand instances by the end of the decade, making it clear that the analytic platform has tapped into a significant market need. They have added hundreds of millions of dollars per year to the billions already being spent with the early entrants—and taken share from incumbent "classic data warehouse relational database management system" products.

Analytic platforms provide two key functions: they manage stored data and execute analytic programs against it. We describe them as follows:

> *An analytic platform is an integrated and complete solution for managing data and generating business analytics from that data, which offers price/performance and time to value superior to non-specialized offerings. This solution may be delivered as an appliance (software-only, packaged hardware and software, virtual image), and/or in a cloud-based software-as-a-service (SaaS) form."*

Some survey respondents, when confronted with this definition, disagreed with it—they consider the "platform" to be the tools they use to perform the analysis. This may be a legacy of client-server days, when analysis was performed outside the database on "rich client" software on desktops. But the increasing requirement for the ADBMS to power the analysis is upending this thinking, and most agreed with our description. We found:

- **The pace of adoption is strong and accelerating**. In 2009, thousands of analytic platforms were sold. And 10 or more players with growing sales are competing for an increasing number of use cases, worldwide, in many industries.
- **The promises being made are being met**. Adopters of analytic platforms report that they tested difficult problems in proof-of-concept (POC) exercises, and the selected products were equal to the tasks—beating their incumbent DBMSs.
- **The right selection process is essential.** Successful POCs require an understanding of the likely analytical workloads—data types and volumes, the nature of the analysis, and the numbers of users likely to be on the system. And real tests separate winners from losers: often, some candidates can't get it done at all.

Introduction

We conducted an online survey of several hundred professionals worldwide, who shared their experiences and opinions with us. Survey results are shown at the end of this report; we include some highlights throughout. Only 25% of those surveyed said they have no plans for an analytic platform. 44% said they are already using one (see Figure 1).

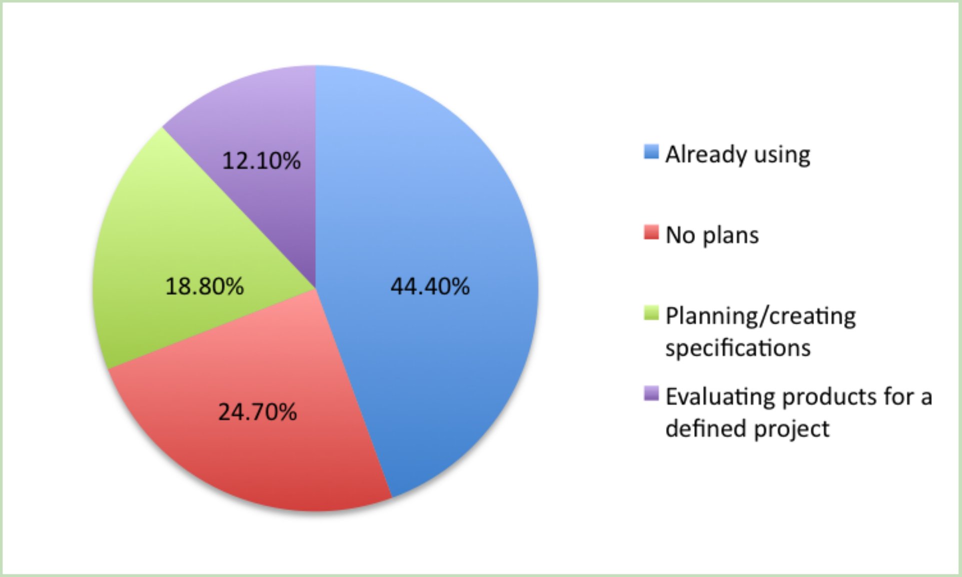

Figure 1: Are You Using or Planning to Use an Analytic Platform?

We also conducted interviews with 8 analytic platform vendors, all of whom are targeting this market, and with a nominated customer from each. The interviewees are quite different from the overall survey population. While our survey showed organizations using database management system (DBMS) products for analytic platforms in proportions that mirrored overall market shares, our interviewees come from the leading edge of the disruptive analytic platform phenomenon. They work for organizations that continue to use classic relational database management systems (RDBMSs) for many applications, including some of the business analytics being targeted by the vendors of analytic platforms, but have opted to use specialty platforms for a variety of reasons.

What we learned was profound; businesses, more and more driven by their need for analytic processing of enormous amounts of data, are responding to the emergence of a class of DBMS specialized for analytics, recently introduced to the market in most cases. A thousand sales of these products in just a few years, generating billions of dollars in revenue, herald the arrival of the analytic platform as a category to be watched closely. It solves important problems, and customers are deriving enormous value from it, creating new classes of business applications and driving top-line growth.

Our interviewees were unanimous: their money was well spent, and their existing classic RDBMS offerings fell short. By contrast, only 21.4% of 168 survey respondents, many still using classic RDBMS products for their analytic platforms, pronounced themselves fully satisfied with their analytic platform projects. While we did not ask for their reasons for this dissatisfaction, some can be derived from the "issues that led you to add an analytic platform" data: the need for complex analyses, query performance, and on-demand capacity topped the list. These issues are mirrored in the case study interviewees.

This report examines the analytic platform, the business needs it meets, the technologies that drive it, and the uses analytic platforms are being put to. It concludes with some guidance on making the right choices and getting started with the products of choice.

The Business Case for Analytic Platforms

What is an Analytic Platform?

Informally: the analytic platform is a response to the inadequacy of classic RDBMSs for new, more sophisticated demands for business analytics. It combines the tools for creating analyses with an engine to execute them, a DBMS to keep and manage them for ongoing use, and mechanisms for acquiring and preparing data that is not already stored. In this report, we focus on the DBMS component of the platform. As noted below, separate providers also offer data sourcing and integration and tools for analytics surrounding the DBMS; these will interact with the DBMS itself and often depend on it for execution.

Why Do We Need Analytic Platforms?

A brief history demonstrates how we got here over several decades. The earliest computing used a simple paradigm for simple analytic processing: business data created by transaction, manufacturing, or other processes was stored in files. Specialists wrote programs to run against them, generating management reports about the state of the business. But routine, multiple, simultaneous use of the data—transactional and reporting—quickly became the expectation.

DBMSs emerged: persistent data stores for many kinds of batch programs to run against—to add, update and delete, and report on data. Online computing made it possible to do these things in real time, and to do them at the same time—multiuser multiprogramming. The client-server era shifted things to a two-or-more-tier model, in which the analytic processing was done on data extracted to a different platform, supporting one or many users working with local copies of the data that might themselves be saved or might go away when the session was done. But this created uncoordinated, redundant, and sometimes conflicting versions of the data.

The data warehouse was envisioned as a central data store where access, definitions, governance, policy, and currency could be centrally managed. Diverse data sources were harvested and data was copied in, separating analytics and reporting from other business processing. Over time, satellite data marts for specific subject areas or user populations or both emerged—along with rising budget authority in business units who desired autonomy. "In front" of these systems, data extraction and transformation products managed feeding the data in; "behind" them, analytic tools for ad hoc reporting, statistical analysis, model building, predictive analysis, data visualization, etc. were created for business users, programmers, and non-programmers alike, to use (see Figure 2). But the DBMS product in the middle of all this was usually the same one in use for everything else.[1]

[1] Teradata, 4GLs like FOCUS and SAS and other products in the 80s were positioned as "storage plus analysis" vehicles for large volumes of data. But most buyers considered their standard, classic RDBMS as the default.

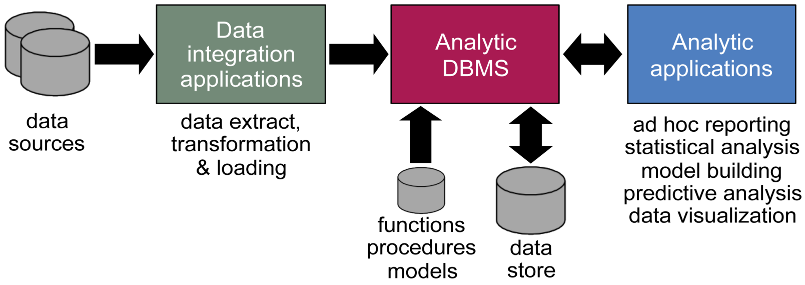

Figure 2: Components of an Analytic Platform

Requirements continued to become more difficult to meet. Online analytic processing (OLAP) added multidimensional capability, the ANSI/ISO SQL standards steadily added more power to the language used in databases, and the TPC-H benchmark was created to measure analytic performance. The benchmark made it clear that DBMSs were coming up short; new approaches were needed, and new vendors emerged to meet them, creating new products that succeeded where the incumbents could not. The forces driving the need for change are largely the same and drove the design of the newcomers who joined the pioneering offerings from Sybase and Teradata.

Data Growth and New Types of Data

The largest data warehouses are now measured in petabytes. Terabytes are not at all unusual, and it's routinely reported that the largest are growing at an increasing rate—tripling in size every 2 years. Sixteen percent of the analytic platforms reported in our survey were managing more than 10 terabytes after loading, tuning, enhancing, and compressing it, and 64.3% said that support for more than 100 terabytes was very or somewhat important in their planning or acquisition.

And while data warehouse data is mostly structured, a significant amount of other corporate data is not. Unstructured text data, weblogs, scientific feeds, photographs, videos, sound files are all potential sources for analysis. Our survey respondents are feeding their analytic platforms with a variety of these new data types: 44.6% are using XML data, 23.8% unstructured file data, 23.2% weblogs, and others (see Figure 3). But the languages, analytic products, and storage mechanisms that have been the everyday toolkit for business analysts were not designed for these new forms of information and often are not well-equipped to work with them.

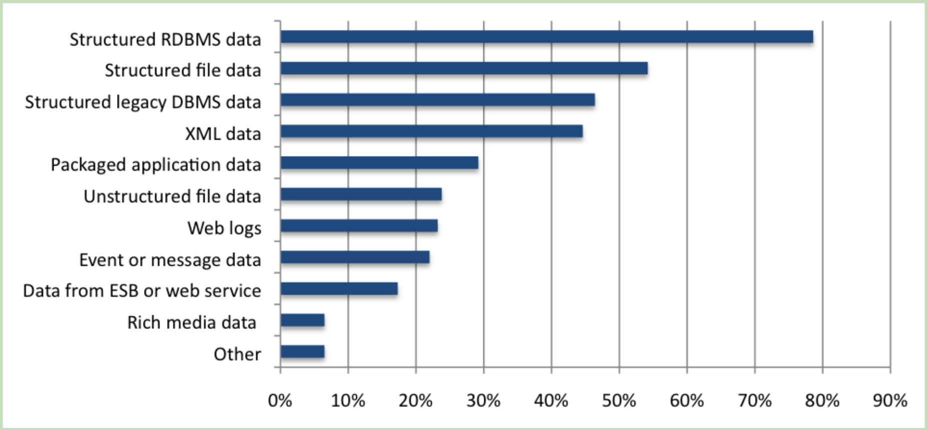

Figure 3: What Data Sources are Used in Your Analytic Platforms?

Analytic platforms are designed to manage large data volumes, sophisticated analytics, and newer data types. They use modern storage paradigms that allow retrieval by columns more efficiently and encode data for better compression. Some use "smart storage" to do some of the work at the storage layer to free up the processor for the heavy analytic lifting. They lash together many commodity processors, with larger memory spaces. They connect processors with one another and with data storage across faster networks to scale processing and storage in sync. They are designed to handle new types of data, even user-defined types that may have specialized code associated with them to make sense of the unfamiliar content. Classic RDBMS products were not built with these innovations in mind and are not always easy to update to leverage these new opportunities.

Advanced Analytics

Simple reporting, spreadsheets, and even fairly sophisticated drill-down analysis have become commonplace expectations and are not considered "advanced." While the term is frequently debated, it's clear that even "simple" analysis is advanced when it needs to performed on a massive scale. Even a simple comparison of profitability across 2 days' trade activity for the top 10 traders each day, for example, is a performance challenge for many systems when run against today's extraordinary volumes of data while other activities run on the same system.

But increasingly, the nature of the analysis itself is more "advanced." Sophisticated statistical work is becoming commonplace for market basket analysis in retail, behavioral analysis in clickstreams for websites, or risk analysis for trading desks. Building predictive models and running them against real-time data is a frequent use case. Some firms require geographic visualizations of data, often against variable shapes such as sales territories or watersheds that are not easily computed. Such ambitions used to be left to the largest firms with highly sophisticated programmers and expensive hardware in their data centers. No longer; savvy business leaders, even in mid-size firms, expect the same from their teams today. And they are doing it outside their classic RDBMS; in our survey, 53% of 223 respondents said they perform business analysis on data not contained within an RDBMS. Nearly two-thirds of them were using hand-coded programs as opposed to packaged tools.

Analytic platforms address the specialization implicit in handling analytic workloads. They retrieve and manipulate large sets, using the right subsets of the fields in individual records. They support large memory spaces for this processing, dramatically improved I/O times to get the data there from storage, and support not only advanced SQL's capabilities but also user-defined functions (UDFs) and programming languages that analysts and statisticians often use instead of SQL. And they leverage new paradigms like MapReduce programs, which may run over external files or against data imported from those sources.

Scalability and Performance

Data scalability is only one dimension—the other is multiuser performance. It has long been a goal of business intelligence (BI) thinkers and planners to involve more users in the corporate analysis of performance. In the client-server era this was often handled by putting tools on their desktops and moving data to them, creating coordination problems as computational models were duplicated. Unsynchronized, often contradictory analysis resulted.

Centralizing the key metrics and algorithms, and making them consumable by more employees and partners who can collaborate around their work, are key challenges. Our survey users expect high volumes of simultaneous usage—32.1% say they need to support more than 100 concurrent users.

Analytic platforms are designed to leverage higher bandwidth connections across a fabric of processors. They utilize modern "container" constructs in memory, used to protect and coordinate multiple processes running in massively parallel scale-out architectures with more processors. They use inexpensive hardware that can be added without taking systems down, so as demands scale, so can processing power. They are designed to cooperate with virtualization layers in modern environments that permit the elastic setup and teardown of "sandboxes" where new analyses and ideas can be tested. All of these capabilities permit analytic platforms to raise the performance profile.

Cost and Ease of Operation

As data volumes, analytic complexity, and the numbers of users all grow, so does cost. Even "commoditized" hardware costs millions of dollars; capital costs expand with data, power, and populations. Power, cooling, space, and backup/recovery for all of it add more expense. Moreover, additional disks and more processors mean more management. Policies across multiple classes of users, security management, and the need to manage environments that cannot be taken down for maintenance all create their own demands and costs.

The number of moving parts in these systems creates its own added challenge: the difficulty of "standing the system up" in the first place becomes an exercise in coordinating software versions, device drivers, and operating systems. Each piece of a complex stack of software is frequently updated by its supplier—and one piece's fix breaks another piece. Systems management skills become expensive, and budget is consumed merely "keeping the lights on."

Analytic platforms offer multiple deployment options that can reduce many of these costs. As they generally move to commodity hardware, some of the pricing premium in older proprietary systems is eroded. The replaceable form factor of massively parallel processing (MPP) systems makes scaling smoother and more granular. It is simpler to add blades with processor, memory, and storage that snap into racks and can be bought as needed. Open source software used in many stacks lowers licensing costs.

Appliances—pre-integrated, preconfigured collections of hardware and software or bundles of multiple software parts that may be installed on any commodity hardware system—offer a way to reduce setup cost. They are increasingly maintained and updated by their suppliers in a way that is designed to ensure that changes don't "break things."

Finally, moving the analytic platform off premises in one fashion or another provides the maximum reduction in cost of ownership and operation. Several vendors will host the system and the data as a dedicated facility. Some will make it available "in the cloud" in a multi-tenant fashion, where tools are shared but data is stored and managed for individual customers. They may take over the process of importing the data from its source systems, such as retail or online gaming systems, and provide the data integration as well as the storage and analytics.

Recall our formal definition:

> *An analytic platform is an integrated and complete solution for managing data and generating business analytics from that data, which offers price/performance and time to value superior to non-specialized offerings. This solution may be delivered as an appliance (software-only, packaged hardware and software, virtual image), and/or in a cloud-based SaaS form.*

In this report, we consider DBMS offerings that form the heart of the analytic platform.

Types of Analytic Platforms

For the past few decades, RDBMS products have formed the data management underpinnings of a wide range of both transaction and analytic IT applications. Products that target analytic processing can be thought of as ADBMSs. Some ADBMS products support SQL and the relational model, while others offer alternative languages and data models.

There are numerous features and functions that differentiate ADBMSs from one another, but for the purposes of simply describing the players, they may be classified in several key dimensions:

- **Use of proprietary hardware**: Some vendors create their own specialized hardware to optimize processing. Others run on any standard hardware.
- **Hardware sharing model for processing and data:** Increasingly, ADBMS vendors support MPP architectures which distribute processing across many blades using chips with multiple cores and significant amounts of dedicated on-board memory. These may have dedicated storage in a shared-nothing environment or may be connected to shared storage such as a storage area network (SAN).
- **Storage format and "smart data management:"** Many ADBMSs are using columnar storage, which dramatically improves the performance of disk I/O for certain read operations. Some support both row and column format in one hybrid form or another. Some also add intelligence at the storage layer to pre-process some retrieval operations. All use a variety of encoding, compression and distribution strategies.
- **SQL support**: Support for "standard" SQL tends to depend on which standard you mean; no vendor supports all of the SQL languages. The absence of some specific features, like correlated subqueries or joins across tables on separate nodes, can be a serious performance problem, preventing some queries from running adequately or at all.
- **NoSQL too.** Recently, a number of offerings have emerged for analyzing specific data types such as documents, unstructured text, and other content not typically stored inside RDBMSs. These are often collectively referred to as NoSQL solutions. Some actually store data while others, such as MapReduce, may operate on files stored in a file system like the open source Apache Hadoop Distributed File System. These offerings are relatively specialized at this time, but can be very effective. Many are adding more features that provide data import, SQL query, and other RDBMS-like functionality.
- **Programming extensibility**: ADBMS engines offer varying degrees of support for the installation of functions, usually described as UDFs, which offer callable computations and data manipulations that are difficult to reproduce with standard SQL. Some offer libraries of such functions themselves and with partners, and some of these take advantage of system parallelism for performance improvement.
- **Deployment models.** ADBMSs may be delivered as an appliance: a complete package of hardware and software; software-only products may be deployed on premises on commodity hardware, hosted off premises, or even in public clouds such as Amazon's EC2.

Hardware Directions

Things are changing fast. Several key elements of the hardware mix are undergoing enormous change, with profound implications for system design and its impact on analytic performance.

Memory is the new disk; disk is the new tape. Reading and analyzing data is made much easier when the data all fits in memory; disk I/O problems, the management of buffers, writing out to disk when new data needs to be brought in—all of these become less of a performance challenge. Memory prices continue to drop and the use of solid state disks (SSDs) and flash memory are rewriting the rules. The first all-memory systems are already appearing, and more will come.

More cores, more threads, yield more processing power. The addition of more cores (and processing threads) to chips has similar implications. As software smart enough to break up and distribute the work (parallelization) is given more threads to work with, performance can scale simply with the addition of more standard blades to a system. In MPP systems where storage is dedicated to the processor, this scalability extends not just to power or number of users but also to data volume.

Infiniband and other network interconnects drive speed. The speed of interconnects can be an enormous bottleneck for system performance. Moving data around inside large systems or from one system to another becomes more difficult with larger volumes. Infiniband's raw speed and ability to provide parallel data movement will be a key asset for vendors that utilize it.

Message from the Market: It's Time

Markets change rapidly, but the effects are often not felt for years. The value of already installed software in most categories is several orders of magnitude larger than the spending on it in any given year or two. Maintenance and support costs for installed software dwarfs new spending. But at the leading edge, players and industry analysts are dazzled by new products and new sales.

Analytic platforms are no exception to this. From the mid-1990s to the mid-2000s, Sybase and Teradata were largely alone in the specialty analytic database market. By 2010, they had some 6,000 installations of their products between them. A dozen or so newer vendors, emerging throughout the last decade, added another thousand or so. The several hundred million dollars spent with these newcomers represented the most significant spending shift in database systems in decades.

But in context, these numbers are hardly a blip on the radar. There are hundreds of thousands of DBMSs installed; so-called data warehouse DBMS sales are estimated at $7 billion per year. The ADBMS is in the hands of early adopters, not mainstream customers—even when they are being used by the world's largest enterprises, their use is confined to a business unit, a division, or a team of specialists. Leaving aside Teradata and Sybase, ADBMS vendors collectively generate a few hundred million dollars annually—less than 5% of the data warehouse DBMS market. Small wonder, then, that our survey respondents told us that they typically begin their search for a platform with their incumbent DBMS vendor.

We learned in our interviews that those adopting analytic platforms are agents of change. They are creating new value, new business opportunities, and new customer opportunities. From a competitive point of view, organizations that have not yet assessed ways to leverage these platforms are already behind. That's the bad news. The good news? One of the key findings of this report is that if you know your problem, you can start fast. And get value fast. At lower cost than you may have thought possible.

Techniques and Technologies

In this section, we review some of the key techniques and technologies offered by analytic platforms, and offer some suggestions about things to consider when evaluating these solutions.

ADBMS versus a General Purpose RDBMS

An analytic platform consists of three main software components: the data integration software for transforming and loading source data into the platform's database, the database management software for managing that data, and the analytic tools and applications that analyze the data and deliver analytics to users. In a traditional data warehousing environment, these three components are purchased separately and integrated by the customer. A key difference with an analytic platform is that the vendor does the integration and delivers a single package to the customer.

At present, most analytic platform database management is done by RDBMSs. For the past few decades, RDBMS products have formed the data management underpinnings of a wide range of both transaction and analytic IT applications. Given the trend by many companies toward extreme processing at both the transaction and analytic ends of the application processing spectrum, it is becoming more difficult for a general purpose or *classic* RDBMS to support the increasing number of different uses cases and workloads that exist in organizations.

The broadening application processing spectrum is leading to vendors developing database management software that focuses on a narrower subset of that spectrum. In this report, products that target analytic processing are described as ADBMSs.

Even within the ADBMS segment, the ability of any given product to support a specific use case or workload varies. The challenge in selecting an ADBMS is to match the workload to the product. This is especially true in the case of extreme processing and also in business environments with constantly changing requirements. Often the only solution is to run a POC evaluation using real user workloads.

In our study, we focused primarily on ADBMS solutions that support the relational model and SQL. However, a brief discussion on using a non-relational, or NoSQL, approach is also included.

In our survey and customer interviews we asked people about key technology requirements for an analytic platform. Our objective was to determine the characteristics and features of an analytic platform that were most important to organizations. The survey results are shown in Figure 4. Features rated as *very important* by the majority of respondents were: query performance (77%), reduced load times (58%), good administration tools (54%), fault tolerance and high availability (54%), and integration into the existing IT environment (54%). Other features that received high scores were easy scaling and hardware upgrades (45%), support for commercial data integration and BI tools (44%), and in-database processing (34%).

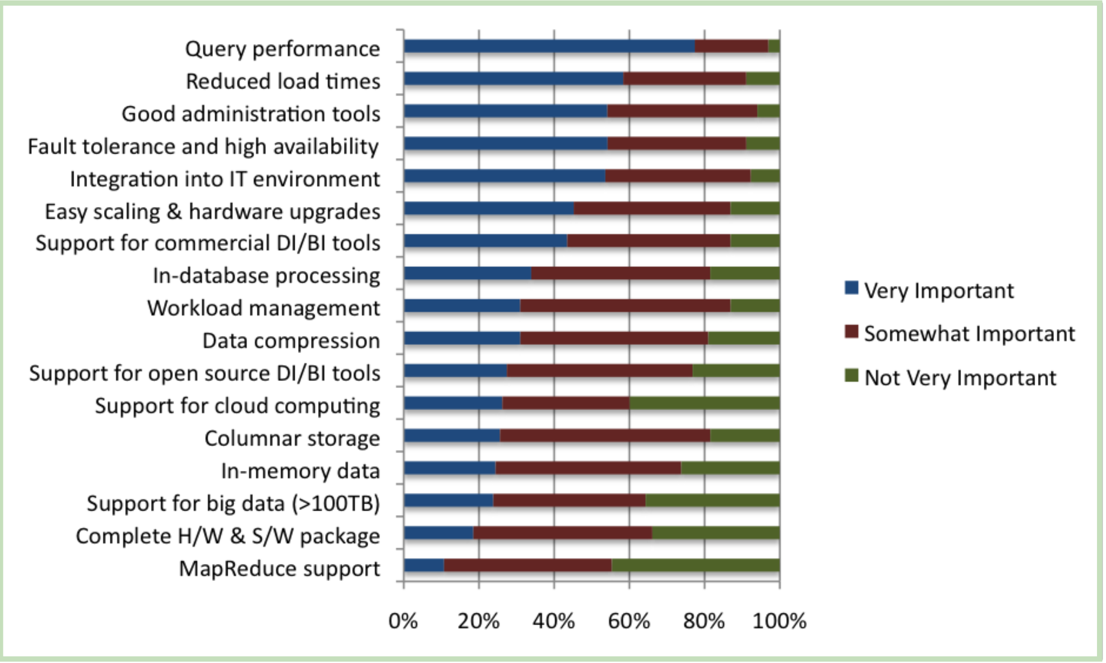

Figure 4: What Features are Important for Your Analytic Platform?

There were no major surprises in these scores except that the score for *in-database processing* was higher than expected. This demonstrates that organizations are beginning to appreciate the benefits of exploiting the power of a parallel database engine to run certain performance critical components of a data integration or analytic application.

Scores that were lower than expected were: support for open source data integration and BI tools (27%), columnar data storage (26%), and a complete hardware and software-packaged solution (18%). The first two results may reflect the limited experience of organizations in using these technologies, while the third score demonstrates that respondents often prefer to have the flexibility to choose their own hardware.

Of the 8 customers interviewed for the report, 7 were doing *extreme processing* involving significant amounts of detailed data and intensive SQL processing. All 7 stated that query and load performance coupled with easy scaling were the main product selection criteria. Most of these customers also required high availability.

ADBMS Application Development Considerations

When RDBMS technology and SQL were introduced in the early 1980s, the big leap forward was separating the user, or logical, view of data from the way it is physically stored and managed. An RDBMS optimizer handles the mapping of SQL requests to the physical storage layer. This explains why the quality of a product's optimizer can play a big role in performance. Even today, this physical data independence remains largely unique to relational technology.

From a development perspective, the factors to consider when selecting an analytic platform and its underlying ADBMS are: its SQL functionality, the programming languages supported, the quality of the relational optimizer, and the physical storage options provided. Some of these factors are of more concern to applications developers than the users of interactive analytic tools. For these latter users, the main consideration is whether the SQL support provided by the ADBMS is sufficient to allow the analytic tool to operate efficiently.

As already noted, most of the customers interviewed for this report were using extreme processing, and in all these cases, a certain percentage of the end users were creating their own ad hoc SQL queries. These queries were often very sophisticated, and the analytic platform's SQL support was a very important selection criterion for these customers. Several customers commented that some of the products they evaluated during product selection had inadequate SQL functionality. Also, with certain products, the physical layout of the database imposes restrictions on the SQL that can be used, which of course is contrary to one of the main tenets of the relational model.

One major area of difference between vendors is their support for SQL built-in functions (scalar, aggregate, string, statistical, etc.), UDFs, stored procedures, and other types of in-database processing such as MapReduce, predictive models, etc. The ability to *push* analytic functions and processing into the ADBMS will usually boost performance and make complex analyses possible from users who have the expertise to use such functions, but not the skills to program them. For many of the customers we interviewed, in-database processing was an important feature when choosing a product. The use of such processing, however, can limit application portability between different ADBMS products because of implementation differences.

It is important to note that just because an ADBMS product supports a particular type of in-database processing, it does not necessary mean this processing is done in parallel. Some of the processing functions may be run in parallel, while others may not. All of them provide more rapid implementation, but the parallelized ones offer superior performance. As an example, not all products support the ability to store and run multiple copies of the same stored procedure on multiple nodes of the configuration.

ADBMS Data Storage Options

ADBMS software supports a wide variety of different data storage options. Examples include: partitioning, indexing, hashing, row-based storage, column-based storage, data compression, in-memory data, etc. Also, some products support a shared-disk architecture, while others use a shared-nothing approach. These options can have a big impact on performance, scalability, and data storage requirements. They also cause considerable discussion between database experts as to which option is the best to use. The current debate about row-based versus column-based storage is a good example here. Often these debates are pointless because different products implement these features in different ways, which makes comparison difficult.

In an ideal world, an ADBMS would support all these various options and allow developers to choose the most appropriate one to use for any given analytic workload. ADBMS products, however, vary in their capabilities. Of course, providing too many alternatives adds complexity to the product, to application deployment, and to database administration. A product could automatically select or recommend the best option, and some products are beginning to support this. In general, however, this is type of feature is difficult to implement successfully given the complexity of today's analytic workloads.

The physical storage options supported by an ADBMS product should be completely transparent to the user's view of the data, i.e., the user should not be forced to code SQL queries to suit the way the data is physically stored. Realistically, in the case of extreme processing, some tuning of SQL queries and the building of indexes and aggregates common in classic RDBMSs may still be necessary to obtain the best performance. This was certainly the case for several of the customers interviewed for the report.

Another option of course is go with a product that provides very little in the way of tuning options and instead employ a brute-force approach of simply installing more hardware to satisfy performance needs. The theory is that hardware today is cheap compared to development and administration costs. This is often the approach used in NoSQL products.

The Role of MapReduce and NoSQL Approaches

No single database model or technology can satisfy the needs of every organization or workload. Despite its success and universal adoption, this is also true for RDBMS implementations. This is why some organizations develop their own tailored solutions to address certain specific application needs.

Google is a good example. Like many other Internet-based organizations, Google has to manage and process massive amounts of data every day. A high percentage of this data is not well structured and does not easily lend itself to being managed or processed by a RDBMS. To solve this problem Google developed its own technology. One important component of this technology is a programming model known as MapReduce.

A landmark paper[2] on MapReduce by Jeffrey Dean and Sanjay Ghemawat of Google states that:

> *"MapReduce is a programming model and an associated implementation for processing and generating large data sets .... Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The runtime system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system."*

MapReduce programs manipulate data records that are formatted as *key/value* pairs. The records are produced from source data by the map program. The *value* field of a data record can contain any type of arbitrary data. Google uses this approach to index large volumes of unstructured data. Note that MapReduce is not a new concept—it is based on the list processing capabilities in programming languages such as LISP (LISt Processing).

The MapReduce programming model has now been implemented in several file and database management systems. Google has integrated it into its BigTable system, which is a proprietary DBMS that uses the Google File System (GFS). It is also a component of the Apache open source Hadoop project, which enables a high-scalable distributed computing system. In the case of Hadoop, MapReduce is deployed on the Hadoop Distributed File System (HDFS).

The MapReduce programming model has also been implemented in a number of ADBMS products. Several of the sponsors of this report provide this capability. This hybrid approach combines the advantages of the MapReduce programming model with the power and integrity of a parallel database engine.

[2] http://labs.google.com/papers/mapreduce.html

The advent of MapReduce has led to the development of a wide variety of solutions that offer alternatives to RDBMS technology. This group of solutions is often referred to as the *NoSQL movement*. These solutions include not only products that support MapReduce processing, but also document and XML data, graph data, etc. Examples of software here include Amazon Dynamo storage system, Apache Cassandra (originally developed by Facebook) and CouchDB projects, MarkLogic, and MongoDB.

The availability of NoSQL software has led to a heated debate about the pros and cons of these solutions vis-à-vis RDBMSs. The NoSQL advocates say that NoSQL solutions are superior to RDBMSs and will ultimately replace them, whereas the RDBMS camp say the NoSQL software lacks integrity and reliability.

The NoSQL debate is reminiscent of the object-relational database wars of the 1980s. The reasons behind them are similar. Programmers prefer lower-level programmatic approaches to accessing and manipulating data, whereas non-programmers prefer higher-level declarative languages such as SQL. The inclusion of MapReduce in ADBMS products offers some of the best of both worlds.

One issue with NoSQL technology is that some software organizations are reinventing the wheel by trying to extend NoSQL software with features that RDBMS vendors have spent many years refining and optimizing. In some cases NoSQL software developers are even adding SQL support. A better solution is to recognize that both technologies have their benefits and to focus instead on making the two coexist together in a hybrid environment.

Many ADBMS and NoSQL solution providers agree that enabling a hybrid environment is what most customers want, and are building connectors between the two technologies. Maybe this is why the website *nosql-database.org* prefers the pragmatic term *Not only SQL* to NoSQL.

MapReduce is particularly attractive for the batch processing of large files of textual data. Seven percent of our survey respondents were using MapReduce with Hadoop. One of the customers interviewed for our study was using a hybrid environment where Hadoop and MapReduce were used for processing textual data, and subsets of this data were then brought into the analytic environment using a software bridge from Hadoop to the ADBMS.

Administration and Support Considerations

Good administration capabilities rated high in our survey results (54% rated it as very important) and customer interviews. Several of the customers interviewed also said that simple administration was an important product selection criterion because they didn’t want to employ “an army of database administrators.” Easy administration was particularly important when designing databases and storage structures, and when adding new hardware.

Several of the customers interviewed also noted that as workloads increased in volume and became more mixed in nature, the workload management capabilities of the ADBMS became more important. Some said they wished they had done a better job of testing out mixed workloads in POC trials.

All of the interviewed customers were happy with the support they received and the working relationship they had with their vendors. Several also commented that the vendor was usually very receptive to adding new features to the analytic platform to meet their needs.

Deployment Models

The deployment options offered by analytic platform vendors vary. Some vendors provide a complete package of hardware and software, while others deliver an integrated pack of software and then let customers deploy it on their own in-house commodity hardware. Some vendors also offer virtual software images that are especially useful during for building and testing prototype applications.

One direction of the analytic platform vendors is to provide cloud-based offerings for deployment in the either the vendor's or a third-party data center or for use on an in-house private cloud. In some cases, the vendor may also install and support a private cloud analytic platform on behalf of the customer.

Ideally, a vendor should support a variety of different deployment options for its analytic platform. This gives customers the flexibility to use the most appropriate environment for any given situation. The customer may opt, for example, to develop and test an application in a public cloud and then deploy the application in house. Other customers may wish to use a hybrid environment where some applications are run in house, while others are deployed in a public cloud depending on performance, cost and data security needs.

Use Cases

Based on prior experience, the survey results and customer interviews from our research study, we can identify three dominant use cases for an analytic platform (see Figure 5):

1. Deploying an enterprise data warehousing environment that supports multiple business areas and enables both intra- and inter-business area analytics.

2. Enabling an independent analytic solution that produces analytics for an individual business area or to satisfy a specific business need

3. Facilitating the filtering, staging, and transforming of multiple data sources and types of data for use in analytic processing

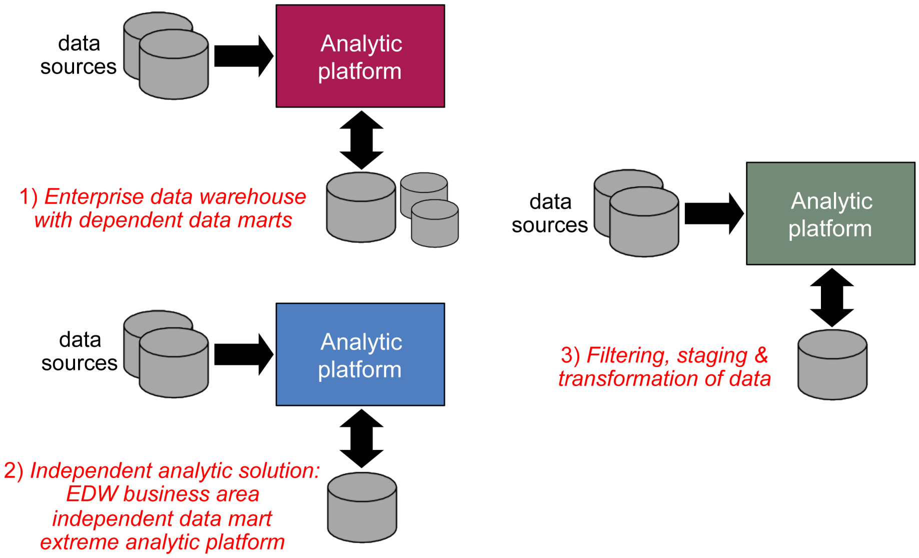
data sources
Analytic platform
1) Enterprise data warehouse with dependent data marts
data sources
Analytic platform
3) Filtering, staging & transformation of data
data sources
Analytic platform
2) Independent analytic solution: EDW business area independent data mart extreme analytic platform

Figure 5: Analytic Platform Use Cases

Before looking at each of these use cases in detail, it is important to comment about the survey results and customer interviews used in this section of the report.

The organizations and users surveyed represent a wide spectrum of industries, data warehousing environments, and technology maturity. The customers interviewed for the report, on the other hand, were recommended by each of the report sponsors, and were, in many cases, developing analytic solutions where it was not practical to maintain the data in a traditional data warehousing environment.

The results and opinions from the two groups therefore sometimes differ. The survey audience results reflect the ongoing evolution of the traditional data warehousing environment, whereas the opinions of the interviewed customers demonstrate the disruptive forces taking place in the industry that enable completely new types of analytic application to be developed.

Use Case 1: Enterprise Data Warehousing

This use case is well established and represents what can be considered to be the *traditional* data warehousing approach. The environment consists of a central enterprise data warehouse (EDW) with one or more virtual or dependent data marts. The data in the EDW and data marts has been cleansed and transformed to confirm to IT designed data models and may be kept almost indefinitely for historical reporting and data analysis purposes by multiple business areas.

The survey results (Figure 6) showed that 68% of survey respondents were using an analytic platform for deploying an EDW, while 42% were using the analytic platform for a dependent data mart containing data extracted from an EDW.

Of the 8 customers interviewed for this report, only one was using an analytic platform for enterprise data warehousing. For this customer, reducing software costs was the main reason for moving to an analytic platform from a classic RDBMS product (i.e., a database system that is used for both transaction and analytic processing).

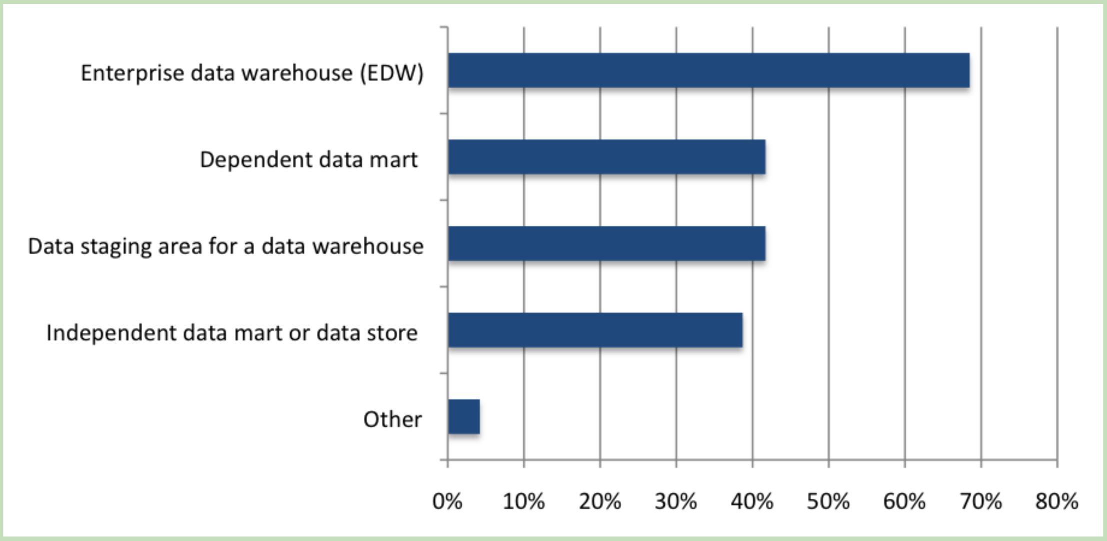

Figure 6: What Use Cases are Being Deployed on Your Analytic Platform?

Use Case 2: Independent Analytic Solution

In this use case, the data being analyzed is maintained outside of the EDW environment. Some 39% of survey respondents were using an analytic platform for this use case. There are three main reasons why an organization may choose to implement this approach:

a) *The organization is deploying analytics and data warehousing for the first time.* In this situation, the analytic platform may be the initial step in building out a traditional EDW environment. One of the 8 customers interviewed for the study fit into this category. The customer chose an analytic platform that enabled the organization to start with a small data warehouse appliance, but grow, via a set of scalable offerings, to provide a large EDW system that can support multiple business areas.

b) *The organization does not have sufficient budget, time, or resources to incorporate the data into an existing EDW*. In this situation, an analytic platform offers the promise of deploying this so-called *independent data mart* solution at a lower cost and a shorter time to value. In the future, depending on business need, the data in the data mart may be integrated into an EDW. Many companies have learned from experience, however, that independent data marts may save time and money in the short term, but may prove more costly in the long term because data marts have a tendency to proliferate, which creates data consistency and data integration issues. As a result, many experts have a negative view of the independent data mart approach.

c) *The organization needs to support extreme processing where it is unnecessary or impractical to incorporate the data into an EDW*. Six of the customers interviewed for this research report match this scenario. Depending on business need, the independent analytic solution may acquire data from an EDW to augment the analyses and may also replicate the processing results back into an EDW. Some independent analytic solutions may be experimental in nature or may only exist for a short period of time to fulfill certain short-term analytic needs.

The first 2 reasons, or scenarios, just outlined are well understood because they are normally a part of the traditional data warehousing life cycle. The extreme processing scenario, however, is relatively new and represents the biggest potential for business growth and exploitation of analytics. It is important, therefore, to look at extreme processing in more detail.

There are several factors driving the need for extreme processing. The first is the growth in data volumes, number of data sources, and types of data. As we noted earlier, many organizations are now generating tens of terabytes of data per day. For these organizations, it is becoming impractical, or even impossible, for cost, performance, or data latency reasons to load certain types of data (high volume web event data, for example) into an EDW. In some cases it may not even be necessary. The application may involve data that only has a useful lifespan of a few days or weeks. Note, however, that these latter types of applications do not preclude the analytic results, or scored or aggregated data from being stored in an EDW for use by other analytic applications.

Another factor driving extreme processing is the nature of the analytical processing itself. BI users are becoming more knowledgeable and more sophisticated in their use of analytics. They want to analyze detailed data as well as aggregated data. They are also building more complex analyses and more advanced predictive models. There is also an increasing demand by these users for enabling ad hoc analyses, in addition to the more traditional predefined reports and analyses provided by IT.

Extreme data coupled with extreme analytical processing leads to the need for high performance and elastic scalability. In data-driven companies, many analytic applications are mission critical, and reliability and high availability are therefore also of great importance. Given constantly changing business requirements and data volumes, the analytic platform in these situations needs to support flexible hardware growth and also be easy to build, manage, expand, and if necessary, tear down and replace. These extreme needs require a new approach to data warehousing, and, in our opinion, this is the sweet spot for new and evolving analytic platforms. These analytic solutions do not replace the traditional data warehousing approach—they extend it by enabling extreme processing.

To use the term *independent data mart* to describe the underlying data store and data management system supporting extreme analytic application processing misrepresents this new breed of applications and the business benefits it can provide. Perhaps a more suitable term would be an *extreme analytic platform*.

Use Case 3: Filtering, Staging, and Transformation of Data

The objective of this use case is to exploit the parallel processing power of the analytic platform's ADBMS to perform data filtering and transformation. This approach is particularly useful in environments involving high volumes of data and/or a wide variety of data sources and types of data. Note that the NoSQL software (Hadoop with MapReduce, for example) discussed earlier is a strong competitor to this approach.

The processing of the data in this use case is typically done using an ELTL approach where the:

- *Extract* step collects and filters source data
- *First load* step loads the filtered data into a set of temporary staging tables in the ADBMS
- *Transform* step does the required transformation and integration of the filtered data
- *Second load* step loads the transformed data into the ADBMS or a remote DBMS for analytic processing

Some 42% of survey respondents stated they were using an analytic platform for the filtering, staging, and transformation of data. One of the customers interviewed for this report was using an ELTL approach with an extreme analytic platform. The business users in this case were able to analyze both the detailed data and the aggregated results from the ELTL processing.

The analytic processing performed in this use case supports data transformation and aggregation, rather than the creation of business analytics. One use of this scenario is to transform less well-structured data into a more usable format. Textual data (web pages, blog pages, unstructured log files, for example) is a strong candidate for this type of transformation.

This use case also offers an alternative to using extreme processing. Instead of loading high-volume detailed data into an extreme analytic platform, an intermediate system is used to filter and/or aggregate the detailed data so that it is practical and cost-effective to load it into an EDW. Of course, the downside of this approach is that information is lost in the filtering and aggregation processing.

We can see from the research study survey results and customer interviews that analytic platforms are being used to support all three of these use cases. In our survey, we also asked organizations what circumstances caused them to move to, or consider, an analytic platform for supporting these use cases. The results are shown in Figure 7.

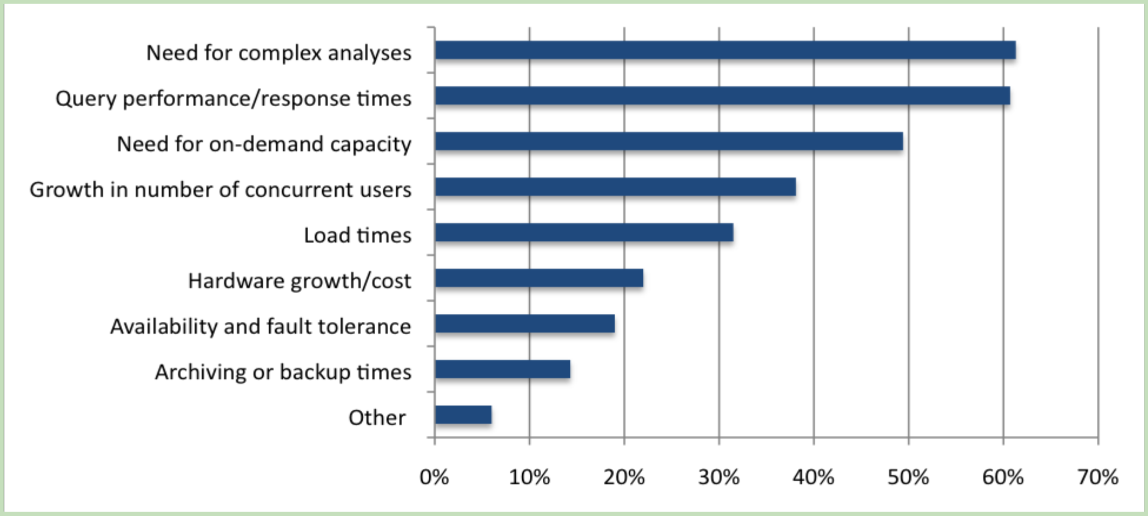

Figure 7: What Issues Led You to Use an Analytic Platform?

The top 5 reasons were: need for complex analyses (61%), query performance (61%), on-demand capacity (49%), growth in user audience (38%), and load times (32%). These results clearly demonstrate that cost is not the main driving force behind using an analytic platform. This was confirmed by our customer interviews. Only one company indicated that reducing costs was the key reason for replacing its existing EDW solution with a new analytic platform.

The top five reasons given in the survey for using an analytic platform, however, do have an indirect relationship to cost. Many of the customers interviewed for this report were building extreme analytic solutions, and the reasons they gave for choosing any given analytic platform all matched one or more of the top five results from the survey. Most of these companies were deploying applications that couldn't be built before. This was either because the application couldn't provide the required performance no matter how much hardware was employed or because the amount of hardware required to achieve acceptable performance was cost prohibitive.

Cost and performance are therefore related, but the key takeaway from the results is that analytic platforms provide cost-effective solutions that extend, rather than replace, the existing data warehousing environment. They enable applications that simply could not be built before.

Getting Started

Success in business is an elusive thing: solving today's question opens new possibilities and new questions for tomorrow. Few categories of technology remain as consistently at the top of CIO planning—"What have you done for me lately?" is the typical question managers are asked. Success for the procurement staff is a signed contract; for business analysts the challenge is greater. The following are some thoughts on how to ensure that the platform selected works today and will continue to grow and evolve as your needs do.

There are also best practices for the implementation of your analytics strategy that leverage the tools you acquire and following them will make success more likely. The vendors and users we interviewed offered some valuable insights and we include them here, together with a quick review of the success factors that make the difference.

Selecting the Right Platform

For analytic platform selection, there is only one place to start: understanding the analytics planned. The use cases in this study hint at the possibilities, but they also make it clear that there is enormous variety: in the skills and preferred tools of the users, the business problems being tackled, the types of analysis required, the latency of the data, and the users' volumes. Will standard reporting be enough? Is ad hoc analysis with drill-down and slice-and-dice operations required? Does internal and external data need to be combined? Will the analyses involve data mining and predictive model building? Will temporary analytic data stores be set up, processed, and then torn down frequently? What are the current and future data volumes and number of users? Know these answers before you begin. Otherwise, making vendor choices is a hit-and-miss process that is likely to lead to project failures.

A vendor POC is the single most vital part of the product selection process. No selection process, however complex, can substitute for the one critical element: testing on your data, with your queries, on the hardware and software platform you plan to use, with the number of concurrent users that matches the expected usage patterns. As you decide who should be on your short list for actual tests, here are some key aspects to consider as you draw up your requirements.

- **Getting the data in and keeping it available**. Ensure that you can load data at the speed you need to absorb it from the sources you expect to use and that any filtering, transformation, and distribution/partitioning you will need to do is supported. Assess the tools offered for design—how complex are they? Do they optimize for your expected queries automatically? Are changes possible without taking the system down for long periods? How does the system provide backup and recovery? How does it assure availability if failures occur?
- **Working with your languages and tools**. Unless you want to train your users and developers extensively, look for products that work with the tools you are familiar with. Consider users beyond the usual suspects internally—you may hope to involve more departments, add more skills, and answer more questions than you have before.
- **Supporting your toughest questions**. If you did your homework, you should know the tough questions that need to be answered and the features that are required to answer them. Complex joins, multipass processing, sophisticated statistics and mashups can make or break products—from the most mature on down.

Other aspects have to do with deployment specifics—don't ignore basics like the interface to your storage hardware, the speed of the interconnects, the level of pre-integration provided across the software stack. Be sure that setting up test or development systems is no more complex than you are comfortable with; analytics is an increasingly iterative process. Explore the possibility of doing such work in the cloud—can the vendor support that? How difficult would it be to move from completed testing in the cloud to production on your hardware?

Finally, the POC trial process is a great indication of the vendor's ability to support you. If the personnel involved don't seem knowledgeable, problems take a long time to resolve, and/or setup seems to take a long time, you must consider what it will be like when the check has been signed and the purchase made. Assess what services are available to you for design, training, and support. And be sure to leave some surprises. **Do not** conduct your trial on prearranged queries and analyses only. Stress test workloads that mirror your expected ones in number of users, volumes of data, and other processes running if there will be any.

Conclusions

Analytic platform adoption is strong, and accelerating. Hundreds of millions of dollars being spent with new vendors represent the most significant spending shift in database systems in a decade. Customers are satisfied; the promises made are being kept. As prospective purchasers test difficult problems in POC exercises, shortlisted products are proving equal to the tasks—beating incumbent classic RDBMSs. The right selection process involves understanding the likely analytical workloads, data volume and types, and numbers of concurrent users—all should be tested. POCs separate winners from losers: often, some candidates can't get it done at all.

Support for open source data integration and BI tools, columnar data storage, and a complete hardware and software-packaged solution are not yet top of mind for purchasers. Conversation with early adopters and survey data show that query performance, support for complex analytics, and on-demand capacity are. As analytic platforms become mainstream, however, it's likely that ease of installation and support and aggressive data compression strategies will begin to grow in importance.

In 2010, analytic platform offerings from DBMS leaders Oracle, Microsoft, and IBM entered the market, and this development should drive increased awareness and growth. The analytic platform will drive billions of dollars in revenue in the next decade, and transform expectations about the ability to use data to improve business results.

Appendix: Detailed Survey Results

Q1: *We define an analytic platform as: "An integrated and complete solution for managing data and generating business analytics from that data, which offers price/performance and time to value superior to non-specialized offerings. This solution may be delivered as an appliance (software-only, packaged hardware and software, virtual image) and/or in a cloud-based software-as-a-service form." Do you agree with this definition?*

Value	Count	Percent %
Yes	209	93.7%
No	14	6.3%
Total Responses		223

Q2: *Are you using or planning to use an analytic platform?*

Value	Count	Percent %
Already using	99	44.4%
No plans	55	24.7%
Planning/creating specifications	42	18.8%
Evaluating products for a defined project	27	12.1%
Total Responses		223

Q3: *How much historical data do you keep online (in non-archival form) for analysis?*

Value	Count	Percent %
More than 3 business years or 12 quarters	95	42.6%
3 business years (or the past 12 complete quarters) or less	55	24.7%
1 business year (or the past 4 complete quarters) or less	51	22.9%
90 days (or the past business quarter) or less	22	9.9%
Total Responses		223

Q4: *Do you routinely perform business analysis on data that is not maintained in an RDBMS?*

Value	Count	Percent %
No	105	47.1%
Yes, with hand-coded programs	86	38.6%
Yes, with packaged tools	32	14.3%
Total Responses		223

Q5: *Which of the following use cases [architectural models] are being deployed for your analytic platform? Check all that apply.*

Value	Count	Percent %
Enterprise data warehouse (EDW)	115	68.5%
Data staging area for a data warehouse	70	41.7%
Dependent data mart	70	41.7%
Independent data mart or data store	65	38.7%
Other	7	4.2%
Total Responses		168

Q6: *Which of the following issues led you to add an analytic platform? Check all that apply.*

Value	Count	Percent %
Need for complex analyses	103	61.3%
Query performance/response times	102	60.7%
Need for on-demand capacity	83	49.4%
Growth in number of concurrent users	64	38.1%
Load times	53	31.5%
Hardware growth/cost	37	22%
Availability and fault tolerance	32	19%
Archiving or backup times	24	14.3%
Other	10	6%
Total Responses		168

Q7: *How much raw data are you managing on your analytic platform? (Raw data is the source data loaded into a data store before adding indexes, aggregate tables, materialized views and/or cubes built from the raw data.)*

Value	Count	Percent %
1 to 10 terabytes	67	39.9%
Less than 1 terabyte	62	36.9%
11 to 20 terabytes	18	10.7%
21 to 100 terabytes	12	7.1%
Greater than 100 terabytes	9	5.4%
Total Responses		168

Q8: *How much data are you managing on your analytic platform after loading, tuning, enhancing, and compressing the raw data?*

Value	Count	Percent %
Less than 1 terabyte	73	43.5%
1 to 10 terabytes	68	40.5%
21 to 100 terabytes	12	7.1%
11 to 20 terabytes	10	6%
Greater than 100 terabytes	5	3%
Total Responses		168

Q9: *How many concurrent users do you need your analytic platform to support?*

Value	Count	Percent %
Less than 20	61	36.3%
21 to 100	53	31.5%
101 to 1,000	38	22.6%
Greater than 1,000	16	9.5%
Total Responses		168

Q10: *What data sources are used to feed your analytic platform? Select all that apply.*

Value	Count	Percent %
Structured RDBMS data	132	78.6%
Structured file data	91	54.2%
Structured legacy DBMS data	78	46.4%
XML data	75	44.6%
Packaged application data	49	29.2%
Unstructured file data	40	23.8%
Weblogs	39	23.2%
Event or message data	37	22%
Data from enterprise service bus or web service	29	17.3%
Rich media data	11	6.5%
Other	11	6.5%
Total Responses		168

Q11: *Please rate the following features that were/are important in acquiring or planning your analytic platform environment?*

	Very Important	Somewhat Important	Not Very Important
Query performance	77.4%	19.6%	3.0%
Reduced load times	58.3%	32.7%	8.9%
Good administration tools	54.2%	39.9%	6.0%
Fault tolerance and high availability	54.2%	36.9%	8.9%
Integration into IT environment	53.6%	38.7%	7.7%
Easy scaling & hardware upgrades	45.2%	41.7%	13.1%
Support for commercial DI/BI tools	43.5%	43.5%	13.1%
In-database processing	33.9%	47.6%	18.5%
Workload management	31.0%	56.0%	13.1%
Data compression	31.0%	50.0%	19.0%
Support for open source DI/BI tools	27.4%	49.4%	23.2%
Support for cloud computing	26.2%	33.9%	39.9%
Columnar storage	25.6%	56.0%	18.5%
In-memory data	24.4%	49.4%	26.2%
Support for big data (>100TB)	23.8%	40.5%	35.7%
Complete H/W & S/W package	18.5%	47.6%	33.9%
MapReduce support	10.7%	44.6%	44.6%

Q12: *Has your analytic platform project met your expectations?*

Value	Count	Percent %
Partially	116	69%
Fully	36	21.4%
No	16	9.5%
Total Responses		168

Q13: *What industry is your company in?*

Value	Count	Percent %
Computer Services/Consulting	35	15.7%
Financial/Banking/Insurance/Real Estate/Legal	32	14.3%
Computer software/hardware/technology manufacturer	22	9.9%
Business Services/Consulting	17	7.6%
Government	16	7.2%
Communications/Telecom Supplier	15	6.7%
Education	15	6.7%
Health/Health Services	13	5.8%
Retail/Wholesale	10	4.5%
Manufacturing/Industry (non-computer related)	9	4%
Other (please specify)	9	4%
Service Provider (ASP, ESP, Web hosting)	4	1.8%
Manufacturing consumer goods	4	1.8%
Travel/Hospitality/Recreation/Entertainment	3	1.3%
Aerospace	3	1.3%
Other	25	11.2%
Total Responses		223

Q14: *How many employees (worldwide) are in your company?*

Value	Count	Percent %
1 to 49	46	20.6%
1,000 to 4,999	35	15.7%
100,000	25	11.2%
5,000 to 9,999	23	10.3%
100 to 249	18	8.1%
10,000 to 24,999	17	7.6%
50,000 to 99,999	13	5.8%
500 to 999	13	5.8%
250 to 499	12	5.4%
25,000 to 49,999	11	4.9%
50 to 99	10	4.5%
Total Responses		223

Q15: *On whose behalf are you completing the survey?*

Value	Count	Percent %
Complete company	73	32.7%
Business department	50	22.4%
Consulting client	39	17.5%
Business division	31	13.9%
Other	30	13.5%
Total Responses		223

Q16: *Please tell us where you and your company are located.*

Value	North America	Europe	Asia/Pacific	Latin America
Where are you located?	59.6%	13.5%	18.4%	8.5%
Where is your corporate HQ?	65.9%	15.2%	13.0%	5.8%

Vertica Overview and Business Description

Vertica (www.vertica.com) led the way in combining two key technologies driving the new generation of analytical platforms: columnar design (storage and execution) and MPP architecture. Years of academic research included the MIT C-Store project driven by co-founder and board member Mike Stonebraker, who pioneered Ingres and Postgres. The company recently brought in a new CEO and several other key executives as it makes the transition from promising startup to a leading player in its space. Privately held, Vertica has grown rapidly, tripling its revenue in 2009 amid a difficult economic environment for IT. With 160+ customers by mid-year 2010, its key markets include financial services, communications services, healthcare, social networking and online gaming, and retail and Web-based knowledge companies. Most sales are direct, but the channel accounts for some 15% of its business, and in 2010 Vertica is launching Asia Pacific operations.

Since its first GA release in Q1 2007, Vertica has delivered two releases per year and is increasing release velocity to deliver three. The Vertica Analytic Platform targets operational, near real-time analytical solutions for high volumes of data. It stresses the value of monetizing data as a function of its time-value (e.g., immediate availability), query performance and broad access by users (e.g., high concurrency). Vertica touts extreme load performance, concurrent and high performance query performance, near-zero admin and elasticity (scaling nodes on the fly). Early wins often resulted from extremely favorable comparisons on total cost of ownership, especially hardware costs; one early win involved a reduction in hardware cost from a $1.4 million Oracle RAC system to $50,000 for a commodity-based platform running the same application.

The Vertica Analytic Database is available as software-only, as a hardware-based appliance, as a virtual appliance on VMware, or online as a cloud computing solution on Amazon EC2. Vertica offers a 30-day free trial version for download to press its case.

Architecture

As its name implies, Vertica was designed from the bottom up as a column-oriented storage and execution platform, with data compression and encoding, and MPP-based architecture. The combination enhances performance dramatically. Making time to deployment a key value drove a focus on automatic database design: Vertica provides a physical design tool that generates and partitions data across nodes based on the input of a logical design, sample data, and sample queries. The output is an automatic physical implementation that requires no manual optimization. These designs automatically account for various sort orders, encoding and compression, and the tool can be rerun to make incremental changes in the background. For example, in a star schema design (Vertica supports any schema type, not just star and snowflake), fact tables will be hash distributed, while dimension tables may be replicated on each node. (In an MPP design, with shared-nothing storage, great benefits are derived from keeping some frequently used data local to the processors, reducing traffic across the bus.) The system also recognizes useful candidates for grouped storage (like storing bid and ask columns in financial services, which are very frequently used together) automatically as well. The automation frees database staff to focus on results, not implementation processes. Other optimizations are automatically generated as well.

Vertica can load 20,000 rows per second per core—which means 240,000 rps on a standard Linux system with two 6 core machines. This scales nearly linearly as you add nodes; customer Zynga (case study follows) loads 60 billion rows per day. Comcast, another early customer, loads nearly a million SNMP message rows per second into a cluster for real-time predictive analytics for network optimization, using standard SQL. Compression operations use multiple, automatically chosen strategies and results vary by data type: Vertica has found compression ratios for telco call data records of 8:1, financial trade execution trails and weblogs of 10:1, and network logs as high as 60:1.

For most Vertica customers, 85% of queries utilize 15% of the columns. Since columns used in predicates are typically sorted, less than 12% of the total compressed data is generally read by a query. Flexstore, added in version 3.5, removes bottlenecks by assigning temporary data, such as intermediate results, to faster storage. It recognizes usage patterns to drive inner versus outer placement on disk. In August 2010, Vertica announced another step—automated support for Flash memory.

Vertica's high availability strategy is based on what is known as k-safety redundancy (replication ensures recoverability from k node failures by storing k+1 replicas). Vertica takes advantage of these replicas by storing the data in different sort orders for further performance improvements. Automated node recovery and shared-nothing architecture eliminate a single point of failure. System administration will soon sport a new user interface for Vertica's enhanced backup and disaster recovery with monitoring graphs for per-user controls for RAM, CPU, and session and runtime quotas. Vertica 4.0 is internationalized via Unicode; its multibyte-aware string functions such as length and substring extend the platform's capabilities to other (non-Roman) languages.

Analytic Functionality

Vertica has steadily added native analytic functionality and today supports many 2-pass, sophisticated SQL-99 functions such as moving window aggregates, advanced time-series analytics for gap filling and interpolation of missing time points (constant or linear), and sessionization. The latter, a frequent use case for MapReduce, applies logic to clickstreams for analysis of client behavior for marketing purposes. Functions are understood by Vertica's optimizer, which takes full advantage of columnar execution versus columnar storage—a key distinction. Statistical functions in Vertica use late materialization—processing much of the data while keeping it encoded and compressed. A U.S. state column can be kept as running counts with only 50 unique state values, coded and compressed. This savings goes beyond the commonly understood I/O reduction in columnar systems based on retrieving only needed columns—memory, communication, and CPU are all conserved and all boost performance.

Vertica supported Hadoop/MapReduce early, but chose not to implement it in the database using UDFs. Customers indicated they didn't want to mix a real-time MPP analytics system with a batch-oriented one; analytic databases tend to be in more continuous use and require low latency. Vertica provides a bidirectional connection to move data back and forth between Hadoop and Vertica for external MapReduce jobs. It plans to release its own UDx framework to bring computation and analytics closer to the data for high performance; APIs will be fully exposed.

Differentiation

Vertica recognizes the differentiation challenge as other analytic platforms begin to compete with it for mindshare. It is engaging around the types of monetization it enables customers to do with their data. Accordingly, it's focusing more on use cases—those that highlight real-time loading of huge data volumes and analyzing it within seconds. These focus on extending the EDW—the data is often not "in there" yet, and some never will be. Vertica is trying to shift conversations to what kind of problem customers want to solve and focus on how fast, at what scale, and at what price. It wins against large incumbent, non-specialty databases, expects continued success against those that provide hardware as a key piece of their value proposition, and is focusing on getting into as many POCs as it can.

Partnerships

Vertica has moved rapidly on the ecosystem front for a young firm. Among its key relationships are agreements with HP (nearly half of Vertica's customers run on HP hardware), Tableau, Red Hat, MicroStrategy, Pentaho, Syncsort, Informatica, IBM Cognos, SAP BusinessObjects, and VMware. Vertica is certified by nearly 30 ETL, monitoring, and BI tool companies. In August 2009, Vertica, Talend, Jaspersoft and RightScale teamed to offer a joint solution stack in the cloud. The solution is now a top 15 template on RightScale, and several customers are in production including online gaming companies such as Sibblingz and CrowdStar.

OEM and reseller partnerships include firms such Unica, NetQoS, Syniverse, and others. Vertica has entered the Federal arena and has begun working with Technica Corporation, a provider of IT solutions for government networks.

How Should Customers Start, and What Matters Most?

Vertica points out that customers should not forget data cleansing and data quality challenges. It believes that the time it saves on database tuning and physical design issues should be spent on these issues instead. It also stresses the importance of doing PoCs correctly; customers should ensure that test data doesn't fit in memory; future workloads likely will not, as data volumes are growing faster than memory improvements. It's also important not to reveal the queries in advance; Vertica finds variability across multiple types of inquiries can be a surprise—and the "typical" query may not be the challenging one.

Future/Road Map Exploitation of Trends

Vertica's expansion will focus on 4 core themes:

- In-database analytics
- Ecosystem integration, especially Hadoop support—leveraging Vertica's footprint reduction to move data around more efficiently
- Elasticity and on-demand analytics—moving data not only between Vertica nodes and adding nodes on the fly, but cloud bursting (private or public cloud) federated clusters of Vertica data (or subsets of data)
- Ease of use

Vertica expects to see increased movement to the cloud. Its experience there dates back to its implementation on Amazon EC2 in 2008 to support PoCs—but customers are using it now for internal clouds. Usage for testing, temporary projects, and workloads will happen with increasing frequency. Effective data compression and flexible physical storage give it a leg up, Vertica believes, and will help with the "Fedexing files" model companies find today as the only reasonable way to deal with high volume data transfers.

Vertica Customer Case Study: Zynga

Company Background

Zynga (www.zynga.com) is an online social gaming business that offers free games for everyone to enjoy. These games range from harvesting plants to making apple pies and playing poker. The company is particularly well known for its Farmville game. Zynga has more than 65 million active users daily and over 235 million active users each month. Zynga game portals exist on Facebook, MySpace, Farmville.com, MSN games, MyYahoo and Tagged. They are also available on the Apple iPhone and iPad.

Zynga is a private company founded in January 2007. The company has raised $219 million in 4 rounds of financing and has some 900 employees worldwide. It is headquartered in San Francisco, California.

For this case study, we interviewed Dan McCaffrey, Director of Data Infrastructure.

The Business Problem

Two years after the company was founded, Zynga still did not have a BI system. Although the company had dozens of games, each game was managed separately. By the beginning of 2009, the company had reached the point in its growth that it needed a BI system to bring together all of the data from its various gaming products for understanding and analyzing its gaming operations. This was seen as a high priority by executive management.

One of the main objectives for building the new BI environment was to discover what motivated customers to play Zynga's games. This information would help product managers to make daily improvements to game content and would also help them in designing new games.

The nature of Zynga's business presented some unique challenges for building and deploying a BI system. Games are accessed from social networking sites and only played for a few minutes at a time. The data volumes involved in tracking and analyzing this casual game play and the social networks involved are very high. Zynga wanted to harvest and analyze this data in real time while games were being played. This would require the loading and analyzing of tens of billions of rows of data per day.

The Analytic Platform Solution

After evaluating several possible solutions, Zynga chose Vertica. During the evaluation process, the company considered several competing analytic platforms and also looked at the possibility of using Hadoop with MapReduce. The main selection criteria were query performance and loading speed. Another important factor was the ability to compress the data to reduce disk storage requirements.

Vertica was selected because it met performance requirements and also achieved significant data compression. Zynga wanted an ANSI SQL-based database approach if possible and chose not to go with Hadoop for this reason. "We took a chance on a SQL solution and it paid off," said Dan McCaffrey. "One thing we learned from the evaluation process is that you really need to know your use cases up front in order to select the right technology."

The Vertica software is installed on two 115-node HP clusters. Zynga plan to increase the size of the hardware configuration to 230-node clusters. "We liked the fact that Vertica uses commodity hardware," said McCaffrey. "It's easy and fast to add new hardware. Only one command is required to get a cluster up and running." He also commented that, "You need to be very selective about what hardware you choose for these kind of environments if you want to achieve high performance and availability."

Detailed data is loaded into the Vertica database in real time using in-house developed software. The raw web data is then aggregated each night using Kettle open source software from Pentaho. McCaffrey noted that, "Vertica's parallelism is ideally suited to this ELT approach because we can push the transformation into the database engine using SQL."

Zynga put together an analyst team that is very experienced in SQL and statistics. Some 80% of reports and analyses are done using custom SQL queries. Software from Tableau is used by less experienced users for producing high-level reports and analyses.

Implementation Considerations

During the initial deployment of Vertica, Zynga did experience some software issues but these were fixed quickly. "The support we received from Vertica was phenomenal," said McCaffrey. "The software issues have now gone away and hardware reliability is now our main concern. To improve availability we have now installed a second hardware cluster."

Following the initial installation, company growth and user adoption of the system caused a dramatic jump in both data and query volumes. Predicting and managing this growth proved to be difficult, and the implementation team became concerned about scaling the system to manage what were likely to be even higher volumes in the future. These scalability concerns caused Zynga to go back and re-run POC trials with several vendors. The results from these trials again led to the decision to use Vertica. So far Vertica is handling the growth, but Zynga is constantly monitoring the system and working closely with Vertica on scalability needs.

The challenge for Zynga is balancing system resources against adoption. "We could limit access to the system," said McCaffrey. "However, rapid adoption is a key success factor for us. The issue is we started off with simple use cases, but we quickly added a significant number of new metrics."

Experience with the new system is making it easier for Zynga to predict growth, but the rapid changes taking place in the social gaming industry still makes this difficult. "You need to predict your scalability requirements and then multiply by 10," joked McCaffrey.

Benefits

The business case for the new system was initially difficult to build, but now that the system is live its value is seen throughout the company. "The return on investment is obvious to us," said McCaffrey. "Metrics from the system impact and enhance every game we produce. The ability to scale was crucial to us being able to develop new metrics to better manage and grow our business. We feel there are few systems out there that can provide this level of scalability. The system is also a crucial underpinning to building new applications."

Summary

Zynga needed an analytic platform that could be used to gather and analyze large volumes of detailed web data about how customers use its gaming software. The Vertica system has enabled Zynga to improve existing products and design new ones that provide customers with the experience they want while at the same time help the company increase revenues. Meeting Zynga's query and data loading performance requirements were key selection criteria, but the ability of the system to scale to meet growth was also a crucial factor. A good working relationship with Vertica and the ability of its field staff to quickly fix problems were also important elements in the success of the project.

About the Authors

Merv Adrian, Principal at IT Market Strategy, has spent 3 decades in the information technology industry. As Senior Vice President at Forrester Research, he was responsible for all of Forrester's technology research for several years, before returning to his roots as an analyst covering the software industry and launching Forrester's well-regarded practice in Analyst Relations. Prior to his Forrester role, Merv was Vice President and Research Manager with responsibility for the West Coast staff at Giga Information Group. Merv focused on facilitating collaborative research among analysts, and served as executive editor of the monthly Research Digest and weekly GigaFlash. He chaired the GigaWorld conference (and later Forrester IT Forum) for several years, and led the jam band, a popular part of those events, as a guitarist and singer.

Colin White is the president of DataBase Associates Inc. and founder of BI Research. As an analyst, educator and writer he is well known for his in-depth knowledge of data management, information integration, and business intelligence technologies and how they can be used for building the smart and agile business. With many years of IT experience, he has consulted for dozens of companies throughout the world and is a frequent speaker at leading IT events. Colin has written numerous articles and papers on deploying new and evolving information technologies for business benefit and is a regular contributor to several leading print- and web-based industry journals. For ten years he was the conference chair of the DCI and Shared Insights Portals, Content Management, and Collaboration conference. He was also the conference director of the DB/EXPO trade show and conference.