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

Table of Contents

Analytic Platforms Research Report

Vendor Overviews and Customer Case Studies

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, Teradata and Sybase IQ, 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

Eight leading analytic platform vendors—Aster Data, EMC/Greenplum, Kognitio, Netezza, ParAccel, Sybase, Teradata, and Vertica—sponsored this report. 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).

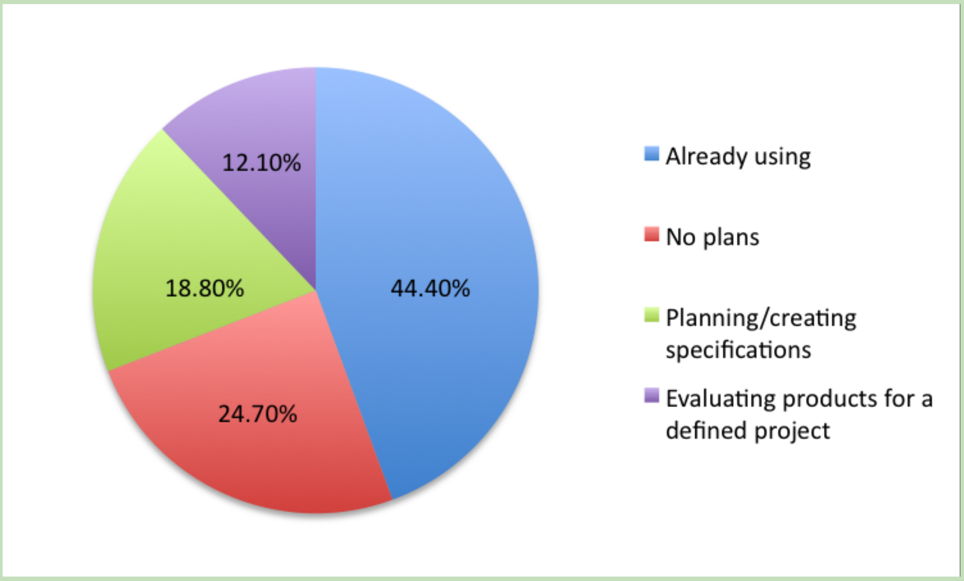

Figure 1: Are You Using or Planning to Use an Analytic Platform?

We also conducted interviews with the vendor sponsors, 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.

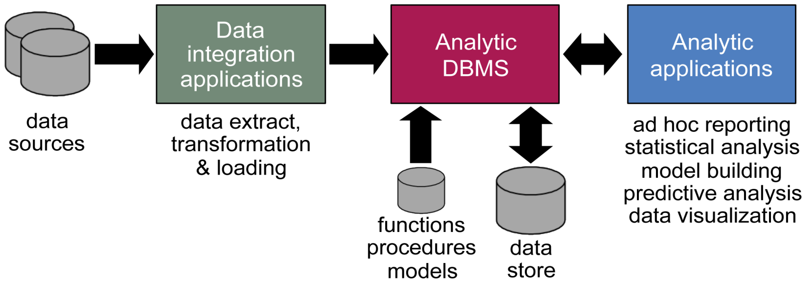

Figure 2: Components of an Analytic Platform

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.

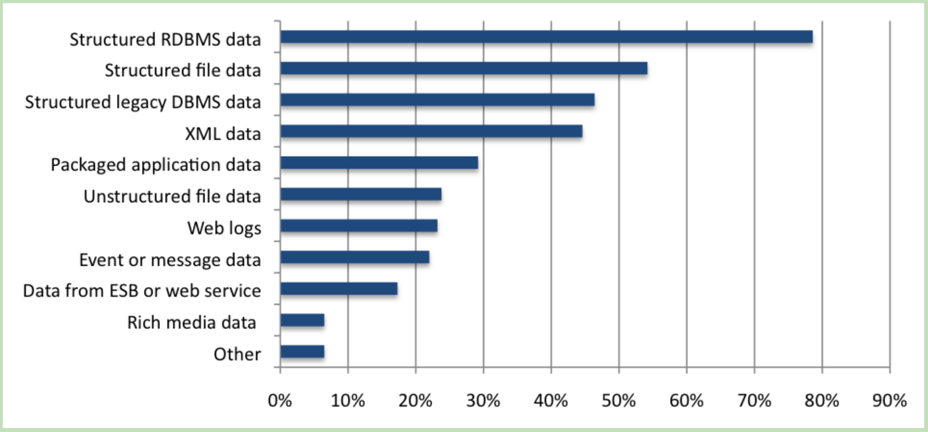

Figure 3: What Data Sources are Used in Your Analytic Platforms?

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: examples in this report are Netezza and Teradata, which still have some proprietary offerings in their portfolio. 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 (some of our sponsors among them), 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 with our sponsors' customers 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%).

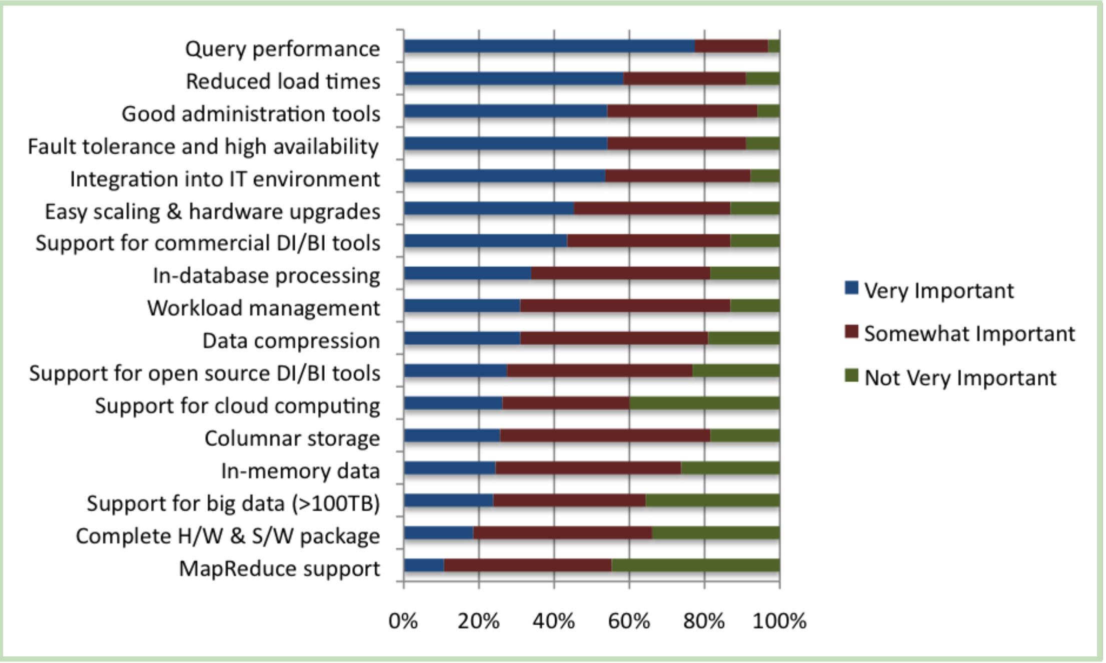

Figure 4: What Features are Important for Your Analytic Platform?

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

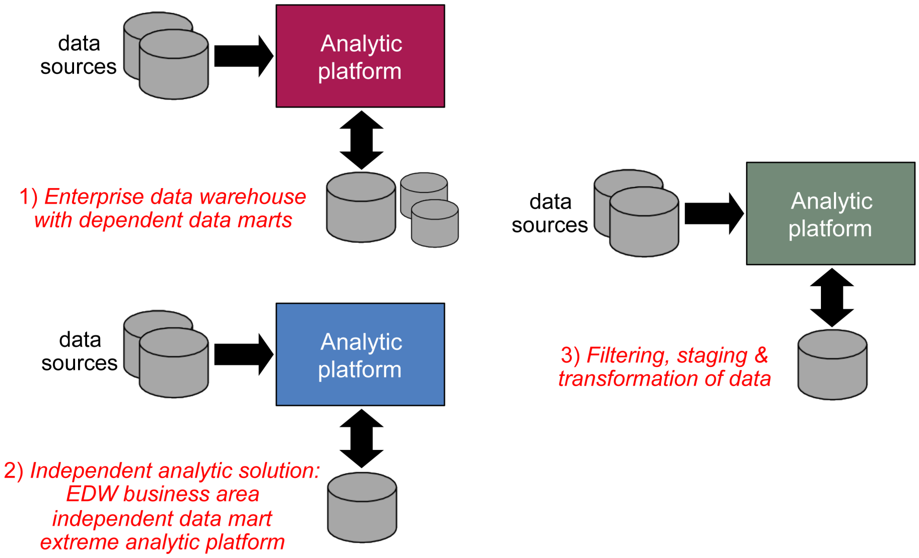
data sources
Analytic platform
1) Enterprise data warehouse with dependent data marts
data sources
Analytic platform
3) Filtering, staging & transformation of data
data sources
Analytic platform
2) Independent analytic solution: EDW business area independent data mart extreme analytic platform

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).

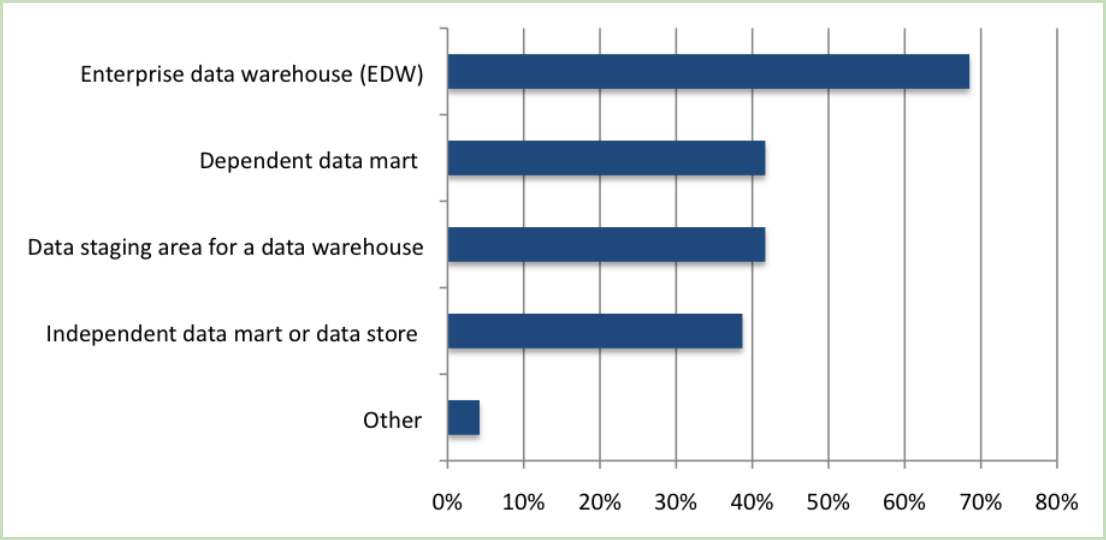

Figure 6: What Use Cases are Being Deployed on Your Analytic Platform?

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.

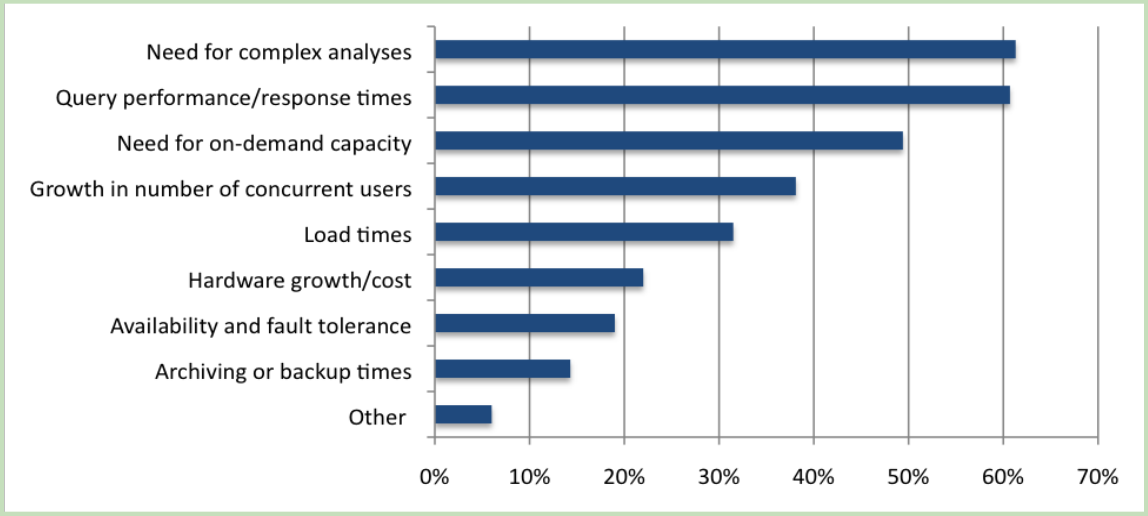

Figure 7: What Issues Led You to Use an Analytic Platform?

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.

The Customer Perspective

We interviewed the representatives of 8 companies that have implemented analytic platforms to improve the effectiveness of business decisions and processes. Each is a customer of one of the sponsors of this report, who nominated them. Our goal was to understand the business case for analytic platforms, how solutions were developed, and the benefits achieved.

This report includes an in-depth case study for each customer. Covered are the specific business problem, the analytic platform solution, implementation considerations, and the benefits achieved. In this section we provide a brief overview of each case study and the benefits customers have achieved. We encourage you to also read the more detailed case studies. They offer valuable information about what it is like to implement an analytic platform.

bet365 (Sponsored by Kognitio)

bet365 is one of the world's leading online gambling groups with over 4 million users in 200 countries. The challenge facing bet365 was how to deploy an environment for producing analytics and optimizing its 24x7 business operations in the highly dynamic world of online gambling. To achieve this goal it needed an analytic processing environment that could provide good performance for the high data volumes and the constantly changing nature of bet365's business. The analytic system had to be able to absorb hundreds of thousands of data records per hour without impacting the timely delivery of the analytics used for optimizing betting operations. Performance was a key issue with bet365's existing Microsoft SQL Server analytic system. Maintaining up-to-date data and querying this data in a timely manner was proving an ever-increasing challenge.

After evaluating several different products and performing two POC trials, bet365 selected Kognitio WX2. The Kognitio platform was selected for several reasons. In addition to satisfying bet365's performance needs and supporting 24x7 operations, it also enabled bet365 to expand and scale the platform to handle expected business growth with minimal impact to business operations.

comScore (Sponsored by Aster Data)

comScore is a well-known provider of digital marketing intelligence. Its services are used by more than 1,600 organizations in over 40 countries. comScore wanted to extend its services beyond those enabled by using data gathered from its 2 million panelists. It wanted to introduce a new census-based service that uses visitor traffic data collected from the web sites of its partners. For these services, comScore needed to be able to load in excess of 18 billion new rows of data a day for analysis. It also needed a configuration that could scale easily as it added new partners to the service.

After running POC trials with several vendors to evaluate performance and functionality, comScore selected Aster Data *n*Cluster. In addition to meeting comScore's performance requirements, other reasons for selecting the Aster Data solution were support for standard SQL syntax, the ability to use MapReduce functions and embed them into the DBMS, and the capability to easily and cost-effectively expand the server and disk configuration to meet data growth.

CoreLogic LoanPerformance (Sponsored by Sybase)

CoreLogic is a key provider of information and services on mortgage financing, servicing, and securitization. Through its LoanPerformance databases, it maintains information on over 100 million active and paid-off mortgages, which are tracked monthly for delinquencies, defaults, prepayments, and foreclosures. The challenge the company faced was that as data volumes grew, the existing SQL Server analytic system could not deliver the required performance. IT staff were spending an increasing amount of time building and optimizing aggregate tables and indexes to try and overcome performance issues.

After running POC trials with several vendors, CoreLogic selected Sybase IQ. Not only did Sybase IQ meet CoreLogic's performance goals, but it also offered the flexibility of supporting commodity hardware. Another key factor in selecting Sybase IQ was its support for the existing MicroStrategy toolset. After the initial installation, CoreLogic saw an 8 times performance improvement and about 40% data compression without making any changes to the analytic applications.

Hoover's Inc. (Sponsored by Teradata)

Hoover's Inc., a D&B Company, is a research organization that supplies information on public and private companies, industries and people. Hoover's Online features an information database covering more than 65 million corporations and organizations, and more than 85 million people. As Hoover's continued to enhance and grow its offerings to add value to the information it offers, the company faced the challenge of not having an integrated view of its customers for sales, marketing, and support processes. Integrated customer information was needed to improve customer retention and to increase revenues by selling new services to existing customers. The company, however, did not have an existing data warehouse and had little data warehousing expertise.

Hoover's decided that the Teradata 2550 data warehouse appliance was the best fit to match its needs. Teradata provided a data warehousing solution that allowed it to start small, but expand as requirements grew. The availability of Teradata Relationship Manager was also important factor for the company. Hoover's deployed the Teradata data warehouse in June of 2009 and achieved a positive return on investment by September 2009.

MediaMath (Sponsored by Netezza)

MediaMath is a leader in the highly competitive and multibillion-dollar display advertising business. The company enables ad agencies and large-scale advertisers to identify, bid on, buy, and optimize ad impressions from a variety of sources. It analyzes upward of 15 billion ad impressions a day and automatically matches each impression in real time with ads that are meaningful and relevant to users. This involves calculating the fair market value of more than 50,000 impressions every second. MediaMath tried a number of other providers to manage this data, but it found that other solutions simply could not keep up with its analytic processing requirements. The company needed a solution that could support the data volumes and fast responsiveness required in online marketing. It also needed a system that was simple to operate and administer.

The system MediaMath chose was the Netezza TwinFin 6. Netezza was chosen after running POC trials with 5 vendors. Only 2 of the 5 could satisfy MediaMath's performance requirements, and only Netezza offered the simplicity and ease of administration the company was looking for.

Large International Banking Group (Sponsored by ParAccel)

This financial institution offers a wide range of services to its over 40 million customers worldwide. One of the bank's trading desks needed an analytic solution that could handle the ad hoc analysis of billions of rows of detailed data about loans and bonds. The solution had to be flexible enough to enable analysts to quickly introduce new queries to reflect changes in business conditions and trends. This was not possible with its existing Microsoft SQL Server system as it took days to test and introduce new queries.

After running POCs with 3 vendors, the trading desk chose ParAccel. Key reasons for selecting ParAccel were its ability to meet the project's performance goals and the flexibility it offered in physical design. The ParAccel system provided significant performance gains in both load and query processing times. Month-end loading is now done in about 2 hours compared to the several days on the Microsoft SQL

Server system. The runtime for one the trading desk's major analytic queries has been reduced to 7 minutes compared with the 3 to 4 days it took before.

Zions Bancorporation (Sponsored by EMC/Greenplum)

Zions Bancorporation is a financial services organization consisting of eight commercial banks. It operates some 500 full-service banking offices in Arizona, California, Colorado, Idaho, Nevada, New Mexico, Oregon, Texas, Utah, and Washington. The challenge the bank faced was that it was too expensive to expand its current Oracle environment to support new analytic processing requirements. Even without the new these new capabilities, predicted data growth indicated that continued use of Oracle was cost prohibitive.

Zions spent a year evaluating various alternatives for the new data warehouse environment and ran several POC trials. Cost was the main factor in choosing a solution. The final selection process came down to a choice between the Greenplum Database and a leading data warehousing vendor. The cost for the initial data warehouse configuration was about the same from both vendors, but with an estimated data growth of one terabyte per year, the Greenplum solution provided a lower cost in the longer term.

Zynga (Sponsored by Vertica)

Zynga is an online social gaming business with more than 65 million active users daily and over 235 million active users each month. Zynga did not have a BI analytic system and the nature of Zynga's business presented some unique deployment challenges to deploying such a system. Games are accessed from social networking sites and only played for a few minutes at a time. Zynga wanted to harvest and analyze this data in real time while games were being played. This required the loading and analyzing of tens of billions of rows of data per day.

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 also wanted an SQL database approach if possible and chose not to go with Hadoop for this reason.

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.

Lessons from Our Case Studies

All of our case study customers ran POC trials; all were happy with the vendors they chose. By contrast, in our survey population, only 21.4% said their analytic platform met their expectations. Case study customers achieved significant performance gains, but also cited capacity and elastic scalability for supporting increasing data and query volumes and deploying new applications.

Six of the 8 customers fit the profile of an extreme processing environment (use case 2c) where the organization wished to rapidly load a significant amount of detailed data and analyze it using sophisticated ad hoc queries. Often both the data and query volumes were unpredictable. Five of the 6 customers already had an existing classic RDBMS analytic processing environment, but the products involved could not handle the performance requirements of the application.

Key requirements for these 6 customers were fast load times, good query performance, and the ability to scale easily. High availability and easy administration were also of high priority. For more complex analyses, most of the users wanted to issue native SQL queries in addition to using standard BI reporting and analytic tools.

The remaining 2 case studies were more traditional data warehousing applications. One of these had an existing EDW environment (use case 1) that was proving to be cost prohibitive for supporting new analytic capabilities. POC trials pointed to a solution that reduced costs while at the same time providing easy and cost-effective scalability.

Both identified a good working relationship with the analytic platform vendor with respect to getting the system configured correctly up front, fixing problems, tuning queries, and adding new product functions as key success factors.

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%

Vendor Overviews and Customer Case Studies

Aster Data Overview and Business Description

Aster Data (www.asterdata.com) provides an MPP database management system with an integrated analytics engine to enable cost-effective management of large data volumes and rich analytics on large data sets. Its unique coupling of SQL with the MapReduce analytics framework enables high performance parallel processing. Much of its R&D has been dedicated to the enablement of high performance, scalable queries and advanced in-database analytic processing in its flagship "Data Analytics Server" called Aster Data *n*Cluster. The company believes that processing full data sets together with application logic on one platform is a key requirement for analytic data warehouses, enabling deeper insights from the data, more precise models, and higher performance for more real-time, mission-critical applications. It sees demand for a single platform for different types of applications; its customers want to put all their data in one place. Seventy percent (70%) of the data it sees is not living in the EDW; Aster Data says its customers are looking for a place to aggregate it. In-database processing eliminates the latency inherent in shipping data across the network to a processing tier. Analysts and developers can access MapReduce through standard SQL because of Aster's SQL-MapReduce framework.

Privately held, Aster Data shipped its first product in 2007 and in the three years since has amassed several dozen customers, mostly in high-end deals targeted at large volumes of data like those at comScore, Barnes & Noble, Akamai, and MySpace. Its key markets include digital media, financial services, government, and retail. Most of its business is in the U.S. and Europe, and some has come through a strong partnership with Dell, including joint marketing, sales, and an appliance-based offering.

Architecture

Aster Data *n*Cluster is an MPP database with a hybrid row- and column-oriented data storage architecture and an integrated analytics engine that runs on commodity hardware. It targets large—terabyte and petabyte scale—databases that ingest many rich data types. *n*Cluster uses four node types: Queen nodes for coordination, Worker nodes for distributed analytical processing, Loader nodes for high throughput data loading, and Backup nodes for massively parallel backup. The MPP architecture makes linear scaling possible. Node specialization for loading is also a critical attribute; Aster Data claims 8 terabyte per hour load rates and also supports trickle loading. Data compression rates are competitive; comScore (customer case study follows) reports achieving 8:1 compression for one of its systems.

Its universal computation layer can source data from row- or column-oriented tables and perform both SQL-based and MapReduce-based analysis. Aster Data *n*Cluster has a unique multitiered architecture which enables task isolation and the ability to scale each tier incrementally as needed to meet workload requirements for query processing, data loading, and backup.

*n*Cluster provides fault tolerance with replication, automatic failover, failure heuristics, and clustered backup to prevent unplanned downtime due to hardware or software failures. Its dynamic workload manager allocates processing and compute resources to in-progress transactions, allowing administrators to change priorities in real time with rule-based prioritization (preadmission control) and dynamic resource allocation and reallocation. The Aster Data Management Console offers GUI-based control of physical architecture, process-level inspection, and resource control.

On premises (appliance or otherwise) is not the only deployment option; Aster Data works with Amazon, AppNexus, and Terremark to offer cloud deployment. Sharmila Mulligan, EVP of worldwide marketing, notes, "Some verticals have strong preferences—in the federal and financial services space, the preference is on premises—often with specific hardware and even specific operating system releases. Web companies, other than the largest, want to go to the cloud for minimal implementation and maintenance costs."

Analytic Functionality

The MPP environment enhances performance dramatically by running analytic functions on each node in the cluster, parallelized across nodes and even across cores. Aster Data's customers have both application and ad hoc needs that require powerful analytics—which can be either difficult to express with only SQL or perform poorly (due to complex joins.). The patent pending SQL-MapReduce framework enables powerful procedural programmability integrated tightly with standard ANSI SQL for business analyst simplicity and tight BI ecosystem fit. Aster Data's focus on ease of development for rich analytic applications is very visible in its Eclipse-based integrated development environment (IDE), Aster Data Developer Express. It automates the creation of packages into which analysts import their Java code, enables local testing on the client, and provides one-click push down of the analytic application to the *n*Cluster servers. The IDE is freely available for download at Aster Data's website, www.asterdata.com.

The SQL-MapReduce analytics framework is designed to offer parallelized execution while preserving isolation from database operations to ensure high availability and high performance. Unlike stored procedures and UDFs, Aster Data's in-database processing is polymorphic, meaning that functions are not tightly bound to only one table or data structure in the database—they are reusable. Users of the functions can be prompted for values at runtime without requiring developers to rewrite the code. For complex analytics, this can greatly reduce time to value as models, segmentations, or other analytic approaches are tested. *n*Cluster scales MapReduce analysis with the data; Aster Data's dynamic workload manager can automatically redistribute and reallocate workloads. Developer Express also speeds analytic application development by automatically generating MapReduce code so the developer does not need to learn MapReduce but rather can leverage existing Java and SQL skills.

Differentiation

Aster Data *n*Cluster is differentiated by both the MPP database platform and the mechanisms for enabling advanced analytics. *n*Cluster's independently tiered architecture ensures that reads, writes, backup and data loading always occur in parallel. Each tier in the architecture can be scaled independently and runs on standard, commodity hardware, so that *n*Cluster can scale with both data and budget. One-click administration is provided for scaling and ensures that *n*Cluster automatically installs and configures new commodity hardware nodes without requiring system downtime. This means "always-on" availability.

System availability is further ensured with background replication processes that store replicated data across servers so that in the event of failures (which will inevitably happen from time to time with commodity hardware), the system automatically activates the replicas, allowing queries to complete as if nothing happened. Dynamic workload management ensures highly predictable performance and guaranteed service levels for complex mixed workloads, including reads, writes, and loads. Fine-grained policy controls allow administrators to reallocate CPU and storage resources based on in-progress transactions to meet usage requirements.

The layer "above" the engine and the interfaces is where Aster Data focuses its differentiation story—on analytics. To speed development of custom and advanced analytic functions, the Developer Express IDE lets developers write in their language of choice, from SQL to Java, C, C++, C#, Perl, Python, .NET, and R. Aster Data also provides Analytic Foundation, a suite of ready-to-use SQL-MapReduce functions intended to accelerate development. The functions are delivered in two ways. "Business analyst-ready functions" provide deep algorithms (linear regression or K-means) and enable them through SQL so that the user can simply set parameters and turn the function loose on the data, across all nodes in the cluster in parallel. The result can operate like a prompted query—add a user interface and prompt the user. "Power-user functions" are more programmer oriented, highly specific building blocks to be put into programs created by coders.

Analytic Foundation, bundled with the enterprise edition of the *n*Cluster, offers functions including statistical analysis, clustering (including K-means),time series, sessionization, graph and market basket analysis, data transformation, Monte Carlo simulation, histograms, linear algebra, geospatial and text processing (including "unpack," which takes weblogs apart for use in analytics).

Partnerships

Aster Data, a young company, has already built strong partnerships with technology vendors such as Dell and HP; analytics application providers such as SAS; BI tool vendors SAP BusinessObjects, IBM Cognos, and MicroStrategy; and data integration players like Informatica and Pentaho. Its reseller relationships with Carahsoft in the federal market and with Amazon for prospects interested in cloud deployments have also been key.

On the analytics front Aster Data partners with SAS, Fuzzy Logix, Cobi Systems for geospatial development; Cloudera and Impetus around integrating Hadoop; and Ermas for in-database SAS and R.

How Should Customers Start, and What Matters Most?

Aster Data believes customers need to focus on several key areas: performance, scalability, richness, and ease of development. Concentrating on one at the expense of the others will limit flexibility going forward. It's important to envision what types of analytics are needed—advanced reporting, operational apps, ad hoc analysis, or highly interactive analysis. It's preferable to point to a wide range rather than just optimize for one type.

Future/Road Map Exploitation of Trends

Aster Data believes that a platform for diverse data types consolidated into a single platform will be in high demand—combining relational, non-relational, and unstructured data with diverse analytic applications that access data directly from multiple storage engines via a SQL and NoSQL interface, including SQL-MapReduce as a universal analytics framework for all data types. It continues to focus on delivering more out-of-the-box, prepackaged analytics modules (e.g., time-series analysis, cluster analysis, graph, statistical analysis, etc.) to ease development of rich analytic applications. To further ease development of rich analytic applications on Aster Data's platform, Developer Express will encompass higher level functionality to move beyond automated glue code creation and code completion to areas such as drag-and-drop code component.

Aster Data Customer Case Study: comScore

Company Background

comScore (www.comscore.com) is a leading provider of digital marketing intelligence. With approximately two million worldwide panelists under continuous measurement, the comScore panel utilizes a sophisticated methodology that is designed to accurately measure people and their behavior in the digital environment. This information enables comScore's clients to better understand, leverage, and profit from the rapidly evolving world of the web-based and mobile computing. Its services are used by more than 1,600 organizations in over 40 countries.

comScore is a public traded company (NASDAQ: SCOR) founded in August 1999. Among the areas of digital marketing that comScore was first to measure are e-commerce (2001), search (2002), ad networks (2005), global Internet audience measurement (2005), online video (2005), widgets (2007), and mobile Internet browsing (2007). It has some 900 employees in U.S., Europe and Asia. The company is headquartered in Reston, VA.

For this case study, we interviewed Will Duckworth, Vice President of Software Engineering.

The Business Problem

In late 2008, comScore wanted to extend its services beyond those enabled by using data gathered from its 2 million panelists. It wanted to introduce a new census-based service that uses visitor traffic data collected from the web sites of its partners. For these services, comScore needed to be able to load in excess of 18 billion new rows of data a day for analysis. It also needed a configuration that could scale easily as it added new partners to the service. comScore decided, therefore, to look for a new analytic solution for managing the census-based data.

The Analytic Platform Solution

After running POC trials with several vendors to evaluate performance and functionality, comScore selected Aster Data *n*Cluster. In addition to meeting comScore's performance requirements, other reasons for selecting the Aster Data solution were support for standard SQL syntax, the ability to use MapReduce functions and embed them into the DBMS, and the capability to easily and cost-effectively expand the server and disk configuration to meet data growth. Will Duckworth noted another reason for choosing Aster Data was that, "The company was easy to interact with at all levels of management and had a similar culture to that of comScore."

The hardware configuration consists of a 70-node Dell server configuration running CentOS Linux with a storage capacity of 7.2 terabytes of data per node. comScore is expanding this configuration to support higher than anticipated data volumes and the ability to store 90 days of history data. In-house developed data integration software is used for the hourly update of the analytic data store, which consumes about 1.1 terabytes of compressed data per day. Transformed and aggregated data from MapReduce processing in comScore's Hadoop environment is also brought into the Aster Data *n*Cluster system. This is done using file transfers at present, but this approach will be replaced in the future by Aster's Hadoop Connector feature.

In-depth data analysis is done using complex SQL queries developed by business analysts, who often use Microsoft Excel to further process the query results. Analysts are primarily interested in exploring detailed data produced during the last 24 hours to 30 days, but data needs to be maintained in the system for 90 days to monitor trends and to enable the next generation of product development.

Although the initial deployment of the system supports business analysts, the user audience is expected to grow to include comScore development and quality assurance analysts as it goes into full production.

Implementation Considerations

The initial 10-node system was installed in early 2009 and was operational less than 48 hours after it was delivered. "It's easy and fast to add additional hardware as we need it," said Duckworth. Aster Data gave the comScore implementers good access to all levels of management, which enabled them to resolve issues quickly and have improvements made to the software to meet their requirements.

Hardware delivery and software certification delays created some initial problems. This shortage was due to using hardware that was not in general release and not certified by Aster Data. "These delays sometimes caused us to suddenly make big changes to the disk configuration, which can lead to a packet storm of data moving around the system as it automatically rebalances itself," said Duckworth. "We are loading data 24 hours day, and we need to load an hour's worth of data in an hour in order to stay on schedule. It's much better to gradually upgrade the system to avoid potential disturbances," he added.

Perhaps one of the most significant challenges was determining what data business analysts required in the analytic store for analysis. "We needed to have good communications with business users about their needs," said Duckworth. "Given the amount of data involved, changes in data structure, such as adding a new data column, can have a major impact on the system in terms of partitioning, indexing, etc. It also affects historical information and historical comparisons. These types of changes have to be carefully planned," Duckworth added. "The Aster platform enables a range of new types of powerful queries to be created. These again need to be carefully thought out and planned for in order to achieve the full benefits that Aster offers."

comScore has upgraded to the latest release of Aster Data *n*Cluster in part because of its new workload management feature. Duckworth said he was impressed with this feature and the other new capabilities in Version 4.5. comScore particularly wants to leverage the power of the enhanced MapReduce functionality.

Benefits

comScore is currently building out its census-based solution and the Aster Data platform. "We are happy with the system and pleased with the lower cost to scale out the configuration," said Duckworth. "As we move into full production we are planning to take full advantage of the benefits that the Aster MapReduce capabilities offer."

Summary

comScore required an analytic platform for a new line of business. Key selection criteria were lower hardware costs, fast loads times, the ability to scale to support increasing data volumes, and good analytic processing performance. In addition to satisfying scalability and performance needs, other reasons that comScore selected Aster Data *n*Cluster were support for complex SQL queries and the ability to program and embed application processing into the DBMS using MapReduce. A good working relationship with the software vendor was also an important factor.

Greenplum Overview and Business Description

EMC's Greenplum (www.greenplum.com) has been one of the fastest growing of the new analytic database management system players. Founded in 2003, it brought its first release to market in 2006 and quickly established itself with an innovative approach designed to compete with Netezza and Teradata for massively scaled data warehouses. An early partnership with Sun Microsystems helped the early ramp, and Greenplum's early embrace of MapReduce for large-scale parallel processing played a role as well. In mid-2010, EMC (an early stealth investor) acquired Greenplum, which had passed the 150 customer mark, and plans to build a new data computing business unit around it.

Part of Greenplum's value proposition has been price; in 2009, it was offering a 100 terabyte configuration on 2 racks for $1.8 million, nearly an order of magnitude lower than some competitive systems. The software can be bought in a perpetual license model or on a subscription basis. In mid-2009, Greenplum added another dimension with its introduction of the Enterprise Data Cloud concept—a notion of flexible provisioning that allows analysts to "spin up" and then "spin down" new instances to perform analyses that might or might not need to be kept after exploratory work was done. That work led to the introduction of Greenplum Chorus in June 2010.

Greenplum engages with its customers directly to help them get data in and extract value from it, to identify business problems it can help then with, and bring back lessons learned. Vice President of Analytics Steve Hillion describes his charter as "building out analytical practices services and tools around the database, to evangelize and promote the use of Greenplum as an analytic platform, and get involved in customer situations and POCs." This is partly billable services work, but Hillion estimates that he spends 20% or more of his efforts just advising prospects and clients in non-billable time.

Architecture

Greenplum is an MPP, shared-nothing database, running on commodity hardware from partners Dell, HP, and IBM. Like other MPP systems, it uses coordinating nodes and worker nodes. In Greenplum's case, the workers are called Segment Servers; the coordinator is called a Master. In Greenplum's case, separate Master Servers are used for MapReduce processing, permitting simultaneous parallelized processes of both SQL and MapReduce to be running.

Greenplum's distribution of data across servers is automatic, as is its parallelization of work to be done; the parallelization code that Greenplum's engineering team built atop its Postgres base inserts nodes that handle join processing for nonlocal data. The gNet interconnect uses commodity components, making the scale-out simpler and less expensive. Greenplum's "scatter/gather" model of parallelization allows data collection and distribution to scale by adding nodes and uses a pipelined dataflow engine to maximize throughput. Compression on disk is optional and adds capacity while improving I/O performance. Scan and compute are combined and operate together, unlike some competing product architectures. For example, instead of decompressing a block and checking to see if it matches a predicate and then sending it to the compute node, when Greenplum scans a block, it need not be sent to another server for processing; the processing lives with the storage. The only time blocks are shipped to other servers is in the case of a join.

Greenplum's GUI-based performance monitor provides views at the resource level and of individual queries. Detailed information at stages of the query plan can be inspected. Greenplum has contributed enhancements to pgAdmin3, the graphical open source management, development and administration tool for PostgreSQL, and ships and supports it with the database.

Analytic Functionality

Greenplum's support for MapReduce enables functions written in Python, Perl, SQL, C, and Java. Support for compiled C MapReduce map and reduce functions offers optimized binding and data marshaling. Java can be used both for general function declarations and as a MapReduce language. Connection management provides control over how many users are connected and intelligently frees and reacquires idle session resources. Since each machine may be running many queries, dynamic balancing of resources across them also facilitates more effective use.

Greenplum is certified with BI tools including BusinessObjects, MicroStrategy, Informatica, IBM Cognos, and SAS as well as open source reporting and BI packages such as Jaspersoft and Pentaho. SAS can be used directly against Greenplum data using the native ACCESS interface. Similarly, R routines can execute queries against the database via RPostgreSQL (or RODBC). In addition, R functions can be written directly in the database as procedural functions (PL/R). Analysts using Python and Perl also have access to libraries that deliver natural language processing and other useful methods.

Greenplum supports ANSI SQL-03 OLAP extensions. The company is developing a library of advanced analytics functions including logistic regression, the Mann-Whitney U test, the Chi-square test, K-means clustering, bootstrapping, log likelihood measurement, and the conjugate gradient method. Access will be offered through an online repository, the Greenplum Analytics Library, which will include source code and documentation as well as technical support. Some of the functions are offered directly to customers today in engagements, but the library is not yet generally available.

Differentiation

Greenplum's message is undergoing a transition as EMC builds a division around it. But the themes will persist; the company has focused on large companies with large volumes of data that need to generate insights with advanced analytics. The natural constituency includes statisticians building complex predictive models: cutting-edge online businesses with extremely large quantities of data asking questions not asked before related to social networking, geospatial, trends, and anomalies—and often inventing new statistical techniques to work with them.

Equally important is Greenplum's message about flexible provisioning, using private cloud computing techniques and social collaboration for EDW and analytics. It promotes self-service provisioning of data marts and "sandboxes." Discovery and collaboration are promoted through data services for analysts to combine and share useful data sets within and across data marts, build them as needed, and tear them down.

Partnerships

Greenplum's partner program will clearly be enhanced in the wake of its acquisition by EMC; it already included the tool vendors named above. Hardware relationships will be deeper in some cases—and perhaps weakened in more competitive ones; but server vendors, for example, all work closely with EMC. Greenplum's reseller, solution, and consulting relationships will be substantially enhanced and benefit from greater visibility.

How Should Customers Start, and What Matters Most?

Greenplum sees two typical starting scenarios. Its engagements often begin with consulting, creating very tailored solutions. But this is a difficult model to build software around; as Hillion says, "We're not in the business of building analytic apps—we let our customers and partners do that."

At the other end of the scale, Greenplum empowers its customers to "take a random walk through the data and see what you find." Succeeding with this approach requires firms to have a clearly defined set of goals up front. Have a clear end point, but let the data tell you what direction to go in—the Agile development methodology is a good guide. Greenplum sees the Cross Industry Standard Process for Data Mining (CRISP) methodology as a useful approach, but notes that it is orthogonal to the business requirements; use the stages as a guide, not as an end in themselves.

Future/Road Map Exploitation of Trends

Greenplum's planning prior to its acquisition was built around creating an environment where there is a unified approach to analytics at the enterprise level. Leveraging the rapid pace of development with newer languages and applying it to analytics with increased use of unstructured data and more procedural capabilities have been the motion. The intent is to build a cohesive ecosystem around data within a firm and let analysts collaborate and speak a common language.

Analysts work with different data and different techniques from day to day. Greenplum hopes to make it easier to switch from one to another and to move data around to explore new approaches. The benefits of the EMC acquisition will be far-reaching and are not all defined as of this writing. Greenplum's partnership with other parts of the EMC family will dramatically change its ambitions and its ability to realize them, as development technology from SpringSource and storage optimization from EMC itself become part of the plan. Check Hollis, VP of Marketing at EMC, points out that, "Data warehouses contain sensitive information and produce analysis that is either confidential or otherwise privileged." EMC can offer substantial value here as well, including "the seductive appeal of running on-demand business analytics as yet another fully virtualized workload using dynamic resources in a private cloud model. Like running on a good-sized Vblock," Hollis adds.

Greenplum Customer Case Study: Zions Bancorporation

Company Background

Zions Bancorporation (www. zionsbancorporation.com) is a financial services organization consisting of 8 commercial banks. It operates some 500 full-service banking offices in Arizona, California, Colorado, Idaho, Nevada, New Mexico, Oregon, Texas, Utah, and Washington.

Zions Bancorp was founded in 1873 and is a public company (NASDAQ: ZION). It is included in the S&P 500 Index. It has over 10,000 employees and is headquartered in Salt Lake City, Utah. Gross revenue in 2009 was $2.7 billion.

For this case study, we interviewed Clint Johnson, Director of Business Intelligence.

The Business Problem

The IT organization of Zions Bancorporation selects and provides the technology for all 8 commercial banks in the group. The same technology stack is used by all 8 banks and is supported centrally. For its BI and data warehousing environment Zions was using a software platform supplied by Oracle.

Historically, this platform had been used primarily for reporting purposes. The bank wanted, however, to leverage the information in its data warehousing system and expand its use of BI to provide more advanced analytic capabilities. It also wanted to support ad hoc analysis, rather than the fixed reports of the past.

The challenge the bank faced was that it was too expensive to expand its current Oracle environment to support the new requirements. Even without the new BI capabilities, predicted data growth indicated that continued use of Oracle was cost prohibitive. It, therefore, began a search for a new solution that would enable Zions to replace its entire Oracle BI environment (enterprise data warehouse, data marts, and downstream systems) with one that would support its data growth in a cost-effective manner, and one that would also enable it to move to a more modern approach to BI for supporting its business operations.

The Analytic Platform Solution

Zions spent a year evaluating various alternatives for the new data warehouse environment and ran several POC trials. Cost was the main factor in choosing a solution. The final selection process came down to a choice between the Greenplum Database and a leading data warehousing vendor. The cost for the initial data warehouse configuration was about the same from both vendors, but with an estimated data growth of one terabyte per year, the Greenplum solution provided a lower cost in the longer term. Like many customers, Zions experienced much higher data growth than predicted—the data warehouse doubled in size the first year—and so the cost savings from the Greenplum solution were even higher.

The Greenplum 7 terabyte data warehouse is installed on Sun 4540 hardware with EMC disk storage. When asked about the acquisition of Greenplum by EMC, Clint Johnson commented, "I am very pleased about the acquisition, not only because we are an EMC user, but also because it offers Greenplum long-term stability and financial support.

Data transformation and loading is done using a combination of IBM DataStage and SAP BusinessObjects Data Services software. SAP BusinessObjects products such as Crystal Reports handle data reporting and analysis. Zions is also developing in-house BI applications using a Ruby on Rails web development framework.

There are about 500 to 600 users of the data warehousing system who have a wide range of different BI skills. About 10% of these users are power users who write their own SQL queries and build their own reports.

Implementation Considerations

The Zions data warehouse is a hub-and-spoke system consisting of a central enterprise data warehouse and several dependent data marts. The company spent 2009 and the first part of 2010 converting the existing Oracle data warehousing environment to Greenplum. Most of the conversion is now complete, but some data marts still need to be converted. Zions is also building an operational data store.

When asked about implementation issues, Johnson noted, "Workload management was not tested thoroughly enough during the POC. We experienced data concurrency problems right away. This was especially the case with batch reports and power users. These users had to be *fenced off* from the rest of the BI users. As a result, we spent a lot of time tuning the system in the early stages of the project. We expect the new workload management facilities in Greenplum 4.0 to help here."

Zions also installed more servers and doubled main memory. “These upgrades really helped with performance and the cost to upgrade the hardware was small,” said Johnson. “We didn’t have to reconfigure the database when adding the new hardware, which made the upgrade easy to do.”

Benefits

“Measuring the return on investment of the new Greenplum system is difficult,” said Johnson. “We focused initially on cost avoidance rather than ROI. We now have the spare capacity to take advantage of BI to address new business requirements and have the opportunity to build a whole set of applications that were not possible before with Oracle. For example, we have developed a 1.3 billion row sandbox on Greenplum for modeling customer profitability. The results from this analysis are put back into the data warehouse. We are also building an application for analyzing customer churn, which is about 18% to 20% in the banking industry. One model shows, for example, that an online banking capability could potentially reduce churn to 3% to 4% even if customers do not use the feature. We are also building new models to help understand customer behavior. We can experiment with these new models in a sandbox and then put then into the data warehouse for production use.”

Summary

Zions needed a data warehousing environment that supported a hub-and-spoke architecture consisting of a central enterprise data warehouse with underlying dependent data marts. The key selection criterion was an analytic platform that reduced the costs of supporting an existing BI environment while at the same time supporting the development of new BI applications that provided business value to the company. The system chosen was the Greenplum Database.

The Greenplum Database enabled Zions to migrate its Oracle data warehousing system onto a more cost-effective analytic platform. The new data warehousing environment supported Zions existing data integration and BI tools and provided the capacity to expand its use of business intelligence to support the deployment of new applications that were not possible with Oracle. A key success factor in the project was the ability to quickly add new hardware to solve some initial performance and workload issues with minimal impact to the system.

This case study demonstrates that it is important to consider not only the initial cost of deploying an analytic platform, but also the long-term costs of such a system given the significant data growth being experienced in most organizations.

Kognitio Overview and Business Description

Kognitio (www.kognitio.com) is a privately held firm that focuses on helping its customers identify trends and patterns across very large data sets to improve business performance. It was founded in 2005 when a consulting company of the same name merged with WhiteCross, a 1987 firm that offered one of the first MPP database engines, originally built on proprietary hardware. The new company, also named Kognitio, offers the WX2 analytic database management system, which it moved to industry standard hardware in 2005 and began to offer in an appliance footprint (meaning that the firm would preconfigure and sell the hardware and software as a unit.)

In 2008, Kognitio added data warehousing as a service (DaaS) to the portfolio, and today half of its business is delivered in that format, making it quite unique in the marketplace as a provider of all 3 deployment alternatives: hardware, software, and as a service.

Having established WX2 with more then two dozen customers, Kognitio has steadily added both direct and channel sales capabilities and enjoyed its best year in 2009 despite the challenging economy with its first continental European wins, U.S. customers, and footholds in several vertical markets it had not entered previously. Its MPP-based in-memory architecture, the explosive data growth its customers are experiencing, and the increasing affordability of CPUs and memory in blade form are playing to its architectural strengths. Kognitio finds that the demands from business users for quick results without the capital costs and headaches of configuration, deployment, and maintenance are creating growing interest in its analytic platform.

Architecture

The WX2 ADBMS is a mature, enterprise-class offering, although as a software product version 7.1 is relatively new. Its architecture has been MPP from the outset, but the processing focus has always been on in-memory processing, which in its earliest days meant that cost was a marketing challenge. While WhiteCross was able to tackle very large, very computationally intensive problems, few companies had both the interest and the budget to pursue them. But the world has changed, and WX2 7.1, the current version of the product, validates Kognitio's belief that analytics are well served by a specialized platform where a large amount of processing power is brought to the data. In a WX2 system, many CPUs are utilized at close to 100% in order to crunch data faster, via in-memory processing, running atop a Linux kernel with WX2 taking control of clustering and distributing the data.

To the user, the system looks no different than other databases; data is loaded onto disk via parallelized bulk load with minimal design needed. Since indexes do not need to be created, the data becomes available for use immediately. WX2's in-memory model is tuned for high levels of concurrency as well. For Kognitio, it's about throughput; how many analytical operations can you get through in an hour or a day. The product offers workload management options via queuing mechanisms; DBAs can set how many concurrent queries are permitted and direct users or groups to particular queues. The database itself will auto-prioritize short-running queries. This sophistication comes with maturity—concurrency and workload management are often poor when databases are first introduced and get better over time. Kognitio provides an ETL tool, automated scheduling, and numerous other conveniences for its DaaS customers; those using the appliance or software versions may use these or their existing tools for design, ETL, etc.

The MPP design does not depend on a dedicated "head node;" all may be equally involved in any query. Any can be an entry point and take over the distribution of work for the queries it handles. Query execution is subject to optimizer compiler processes; once the system determines which node to use (based on how free it is), that one becomes the head node for that query. WX2 has its own software message passing system written at the UDP level atop Ethernet. Early systems used multiple 1 gigabyte (GB) connections (4 or more); today, Kognitio uses 10 gigabit Ethernet with redundant connections. WX2 divides the traffic; if one fails, it redistributes. The system's ability to scan data scales linearly with the number of processors, since each has its own data automatically distributed on load. Kognitio claims a scan rate of 600 million rows per second per blade; thus a 100-blade system can scan 6 billion rows per second. Fault tolerance, online changes, and hot disk swapping are also supported.

Analytic Functionality

WX2's architecture, being very SQL focused, works well with partnering BI tools such as MicroStrategy, which gets very high marks for its well-formed queries. Kognitio finds that tool users notice the performance benefits immediately, even those now experimenting with solid state disks (SSDs) with their existing non-specialized databases. While SSDs are faster than hard disk, the company says, they deliver nowhere near the performance of true in-memory processing. SSDs are still block access devices, delivering chunks of data containing both useful and useless bits, while in-memory processing provides truly random access. Similarly, the use of big caches, while helpful, requires the system to ask for every record you want, whether it's there or not. In-memory architecture changes the performance of analytics so much that designs calling for materialized cubes, prebuilt aggregates, and so on can be eliminated, allowing analysts to focus on the analytics, not the compensatory data design.

WX2 supports what it calls "plug-ins" and provides several specific ones to industries where it has early-adopter customers. Such frequently seen tasks as IP address mapping and Monte Carlo simulation are available today. Customers may also code their own plug-ins in C and C++. As a result of its services business, Kognitio has also built a number of applications targeted to the industries it has seeded, including suites of web analytics and programs for the analysis of call data records (CDRs) for telcos. Usage reporting can be wedded to predictive analytics, for example, and such uses have been in place for some years with British Telecom (BT), a marquee WX2 user, for pricing models. BT is literally able to do what-if analysis against its entire pricing structure to see the effect of changes it considers.

Differentiation

Kognitio's messaging focuses on two key themes: its architectural differentiation with in-memory processing using an MPP structure on commodity hardware, and its multi-deployment style approach. The former provides flexibility, scalability, performance resiliency, and continuous uptime. Equally important, it lets Kognitio stress concurrency—many users or reports running simultaneously. Loyalty Management Group (LMG) has implemented WX2 as a data warehouse appliance in order to power its application that offers retailers and suppliers the ability to analyze point of sale data to gain insight, intelligence, and statistical confidence. LMG maintains two years' worth of full data, rather than just 10% samples and at times has 400 users online simultaneously.

Perhaps the most unique of all Kognitio's attributes is its ability to deliver as an appliance, software-only for on-premises deployment, or DaaS. Many of its customers use two or three—for example, they may have production on an appliance but maintain their disaster recovery and development in a DaaS as a service. During the POC process, prospects decide—but sometimes they move from one to the other after deployment. Kognitio's business splits roughly 50/50 between DaaS and on premises.

Partnerships

Kognitio is an HP Tier 1 partner; bet365, described in the accompanying case study, is a good example of an HP deployment with rack mounted HP servers deployed in separate configurations for production and development. bet365 began with existing equipment and added an appliance later, finding the convenience of Kognitio's ability to work with its partner for the appliance form factor valuable. Sun, Dell, and IBM systems have been deployed successfully as well with other customers.

Tools vendors MicroStrategy, IBM Cognos, and SAP BusinessObjects are key partners, along with Informatica for data integration and KXEN for predictive analytics based on data mining with a focus on customer life cycle analytics. On the services side, Kognitio has a strong relationship with Capgemini (the two firms recently partnered for a 7-year Royal Mail Group project.) Kognitio has also forged a partnership with Hosted Solutions in the U.S. and 2E2 in the UK for its hosted offerings.

Future/Road Map Exploitation of Trends

Kognitio sees in-memory computing becoming increasingly important, validating its long-held belief in that model. The dramatic price changes and new Intel processors are making the case that memory is more affordable and the performance improvements are demonstrable. The company is not as optimistic about MapReduce, believing that to be moving steadily through its hype cycle and becoming less important as alternatives using full database capabilities demonstrate their ability to handle the kinds of problems MapReduce is being used for.

Kognitio believes another key trend is the increasing requirement for 24x7 availability of full analytic capability against a company's data. While this was not so a few years ago, many parts of the business, not just power users, need accessibility to be there, always available, with no downtime, and geographic challenges for multinationals redouble its importance.

For its own part, the firm sees an opportunity to continue to expand its target markets, noting interest in vertical markets such as automotive insurance and horizontal opportunities like customer service call centers. On the technology side, it plans to offer expanded performance monitoring and connectivity for tools that expect to find cubes; WX2 can respond to those inquiries without pre-constriction, Kognitio asserts, dramatically reducing design, storage, and administration requirements.

Kognitio Customer Case Study: bet365

Company Background

bet365 (www.bet365.com) is one of the world's leading online gambling groups with over 4 million users in 200 countries. The group has more than 1,000 employees and is headquartered in Stoke-on-Trent in the UK. It is the UK's seventh largest private company.

For this case study, we interviewed Martin Davies, Chief Technology Officer of bet365.

The Business Problem

The challenge facing bet365 was how to deploy a BI environment for producing analytics and optimize its 24x7 business operations in the highly dynamic world of online gambling. To achieve this goal, it needed a data warehousing environment that could provide good performance for the high data volumes and the constantly changing nature of bet365's business. During a sporting event, for example, the betting process is very volatile and this can result in the creation of hundred of thousands of data records per hour. These records must be absorbed into the data warehouse without impacting the timely delivery of the analytics used for optimizing betting operations.

Performance was a key issue with bet365's existing Microsoft SQL Server analytic system. Maintaining up-to-date data and querying this data in a timely manner using the existing analytic environment was proving an ever-increasing challenge, and it became obvious this was not a sustainable approach in long term. The company, therefore, decided to look for another solution.

The Analytic Platform Solution

After evaluating several different products and performing two POC trials, bet365 selected Kognitio WX2. The Kognitio platform was selected for several reasons. In addition to satisfying bet365's performance needs and supporting 24x7 operations, it also enabled bet365 to expand and scale the platform to handle expected business growth with minimal impact to business operations. Another reason for choosing Kognitio was the company was easy to deal with, provided good support, and had a similar corporate culture to that of bet365. "Although price is always factor in technology selection, this was not the main driving force in this case," noted Martin Davies.

bet365 uses in-house developed data integration software for updating the 5 terabyte Kognitio WX2 data warehouse. This update is currently done on a daily basis, but more frequent updates are planned in the future to allow closer to real-time analysis. The MicroStrategy BI toolset was chosen for providing data reporting and analysis.

The Kognitio WX2 software is installed on a 20-node HP server configuration with a storage capacity of 1 terabyte of data per node. A 4-node HP server system is used for development purposes. bet365's operational applications remain on its Microsoft SQL Server and Oracle systems.

Several hundred internal users employ the bet365 analytic platform for querying customer information and analyzing the last 3 months of activity to optimize betting operations. Approximately 20 internal users produce more intense analytics on historical data, to compare current business performance with prior years, for example.

Implementation Considerations

The system was developed and deployed in approximately 10 weeks. The biggest challenge facing bet365 during implementation was the complexity of its existing operational database systems and business processes. Betting is a complex business and bet365 supports a wide range of different sports; each one has its own unique operational data structures and business rules.

The company had already developed a set of data transformations and business rules for its Microsoft SQL Server analytic system, but these were not easy to move to the Kognitio WX2 environment. A sound relationship with Kognitio gave the bet365 implementers fast access to all levels of Kognitio management, which enabled them to resolve issues quickly and have improvements made to the Kognitio WX2 software to meet their requirements. Davies commented, "We received good support from Kognitio during the implementation, and they provided everything we needed to meet our objectives."

The system has been in production for about a year. During this time, most of the Microsoft SQL Server BI applications have been migrated to the Kognitio WX2 platform.

Benefits

The Kognitio WX2 data warehouse currently contains 9 years' worth of information, and analysis of this data has enabled bet365 to improve the overall efficiency of its business. To date, the company has not measured the ROI of the Kognitio WX2 analytic platform. "We have not had the opportunity to do this," said Davies. "However, updates are now done in a few hours, instead of days, and queries run orders of magnitude faster. The configuration is also much easier to administer because the Kognitio WX2 software does a lot of the work automatically for you."

Summary

bet365 required a corporate data warehousing environment that could support fast data loading in parallel with good analytic processing performance. The company chose Kognitio WX2 because it could satisfy bet365's performance needs and also easily scale to meet anticipated business growth. Kognitio's WX2's support for 24x7 operations and a good working relationship with the vendor were also important factors in selecting the Kognitio solution.

Netezza Overview and Business Description

Netezza (www.netezza.com) is a pioneer of the new generation of analytical platforms. Its entry into the market in 2001 catalyzed an economic and architectural shift with a dedicated analytical appliance form factor at a dramatically different price point. Netezza is publicly traded and has demonstrated its staying power, adding dozens of new names every quarter—demonstrating the continuing viability of its market positioning. Now with more than 350 customers, its key markets include telco, financial services, government, and retail. Most of its business is direct, although some is now coming through a relatively new distribution partnership with NEC in Asia.

Today Netezza delivers two platforms: TwinFin for large data volumes and Skimmer for smaller ones and for testing. It recently upped the analytic ante by adding i-Class, a library of functions to bring high performance analytics closer to the data. Its appliances are not positioned as generic high-performance compute platforms; from the outset Netezza has focused on data warehouse-specific workloads and the need to scale to large data volumes.

Sampling and aggregation take too much detail away, the company says: to do analytics around the "long tail" of distribution in your data sets, you need more than just a sample. Collecting tens or hundreds of terabytes and running predictive analytics and optimization techniques provides more insight than conventional BI and dashboard reports. In Netezza's view, such efforts will be required to be competitive in the years ahead for many industries.

The majority of Netezza's customers measure their data in terabytes; it describes its "sweet spot" today as being in the 10 to 20 terabytes range. At the low end, a few customers are working with "only" a few hundreds of gigabytes; these are typically departmental or test systems, since the firm does not target SMBs. The largest are now hitting petabyte volumes and asking Netezza to scale there. "Every company we talk to today is collecting more information than they ever have before," says Phil Francisco, VP of Product Management and Marketing. Exploding data types include Internet traffic, call data records (CDRs), SMS and MMS messages.

Architecture

Netezza was the pioneer in data warehouse appliances, beginning with a proprietary hardware model. It uses a hybrid architecture, with a symmetrical multiprocessing (SMP) system as a head node to perform SQL plan building and administration in front of a MPP array of worker nodes. Now in its fourth generation of systems, it has moved to a commodity platform for most of its components. The specialized hardware is in its S-Blades: IBM blades plus a daughtercard sporting a field programmable gate array (FPGA) processor that performs smart storage filtering operations.

Netezza's storage innovations are also based on "Zone Maps," which keep track of key statistics such as the minimum and maximum value of columns in each storage extent. The Zone Map lets the system identify which data falls in the desired data range, often avoiding general table scans and the associated enormous I/O overhead they create. Other vendors are now emulating this demonstrably successful approach. Netezza goes further: in processing the work, the FPGAs make further "smart" decisions; they PROJECT only the columns in the SELECT statement and RESTRICT to retrieve only the rows in the WHERE clause.

Netezza's maturity among analytic platforms is reflected in the relative richness of its feature set, system administration, and workload management features. The effect of all these optimizations is to dramatically improve overall processing speed—and let it scale linearly as more blades are added. Enterprise-class requirements have also driven sophisticated recovery features on S-Blade failure, improved data compression, and query optimization.

Analytic Functionality

Netezza's architecture supports multiple "engines" above an infrastructure that streams data sets efficiently—using FPGA filtering and projecting, but also removing many of the complexities of running on a table-based data warehouse. One powerful example is its support of an engine for matrix manipulation. Many sophisticated mathematical problems, such as matrix multiplication, linear algebra, and others can be programmed without the developer needing to be conscious of the data organization.

The portfolio of specialized engines also includes ones for Hadoop/MapReduce and R. Netezza has an SDK for Eclipse IDE plug-in. It provides wrappers for Java, Python, Fortran, C and C++, and wizards for creating UDFs that builds the surrounding code automatically. For the extensive community using R, Netezza provides a GUI.

One of the most promising features is the growing library of prebuilt functions—some 40 to 50 strong—that will scale to use available memory and will be maximally parallelized. These are callable from any language Netezza supports. Taking it a step further, the vendor offers wrappers around a set of functions from the GNU Scientific library—2,000 of them—also callable from a SQL UDX (Netezza's UDF). Another set of functions in the R community's CRAN repository (which contains 1,900 packages and 4,000 to 5,000 functions) have not been explicitly parallelized by Netezza—they may or may not have been written assuming an MPP platform. Still, they add flexibility and shorten time to delivery for developers.

Differentiation

Netezza is a highly visible and passionate advocate for the notion that analytic engines are better suited for the use cases they target, but it finds that customer decisions are still affected by which DBMS they already use for an enterprise data warehouse (whether they call it an EDW or not). Its calls with financial analysts often include discussions of the competitive landscape with an emphasis on Oracle and Teradata, which it is often displacing. Netezza's success has driven others to focus on appliance form factors, and the entry of Oracle and EMC in particular promises to create significant challenges.

In such an environment, Netezza's large installed base provides a powerful argument that it belongs on short lists, and its increasing focus on analytics is reflected in the stories it tells. Time is a key dimension: time to stand a system up is faster for an appliance. Time to develop is enhanced significantly with the rich set of available and supported functions. And real-time applications—or something very close to them—are growing in importance for its customers. MediaMath, discussed in this report, is all about making decisions in near real time for ad placement on a site, across a number of particular venues. The shops Netezza targets have deep skills in house, developed on analytic tools, not databases. Netezza differentiates itself on a mature system designed to provide flexibility in integrating with a wide array of analytic tools and supporting a variety of frameworks and languages, adding parallel execution and effective management of large data volumes to complete its value proposition.

Partnerships

As one of the more established firms in the analytic platform space, Netezza is farther along than most in building strong partnerships. On the infrastructure software side, the company sports advanced applications for the TwinFin platform developed or in progress from leading firms such as SAP BusinessObjects, Kalido, MicroStrategy, SAS, and TIBCO. A recent agreement with Composite for the Netezza Data Virtualizer will provide federation across multiple Netezza appliances. A partnership with Cloudera will enable data movement and transformation between the TwinFin appliance and Cloudera's Distribution for Hadoop (CDH). On the hardware side, IBM is the key partner as supplier of the blades used in the appliances, and NEC began jointly building systems based on its hardware with Netezza in 2010.

In specific markets, Netezza is partnering with vendors such as Aperio CI, doing interesting churn analysis that permits offers to be made on the spot to customers during support or problem calls. In marketing, analytics to understand how purchasing decisions are shaped—by opinion leaders within a customer base, via segmentation analysis, and with faster visibility into real data—are the basis for partnerships with firms such as Pursway (formerly Datanetis), DemandTec, and QuantiSense. Dozens of large and small systems integrators are working with Netezza, and the firm is building this function out to grow business delivered through its channels above the 20% or so it already has.

How Should Customers Start, and What Matters Most?

Netezza recognizes that analytics still starts with EDWs for most firms. But increasingly, analytics happens elsewhere as well. Francisco says, "As soon as IT represents friction in moving to advanced analytics, the business moves around you, sourcing it themselves." Netezza sees its customers grappling with the challenge of getting people across the business to see the value of doing analytics in one place and set the fragmentation of the analytic function (visible in the growth of MapReduce) aside. Organizations need education to bring business users in—show them how their peers get value and go beyond the CIO's organization if needed.

The key message: get started; don't wait. Netezza likes to refer to a comment from Forrester Research's Jim Kobielus, "Don't let perfect be the enemy of good enough." Exploratory analysts will see values they don't know are there right now. Netezza's customers have been surprised by what they found and expanded from that base. For advanced analytics, it's important to look through the function libraries and understand what is possible.

Future/Road Map Exploitation of Trends

Netezza believes data warehouses will become more operational. Whether they ever get to complex event processing is a question, but shrinking the latency between collection and action is a major trend. Storage technology's movement from rotational to SSD or flash will help make some of this possible. It's happening in early adopter use cases already; Con-way Freight is optimizing the loads in its trucks and says it deals with customer churn just like telcos. To avoid having people who are dissatisfied leaving, it's moving to using real-time analytics to make them an offer if they have a problem as well as proactively sending offers out on an ongoing basis.

Netezza believes that being model-agnostic is important. Enabling its customers to be unconcerned about physical or logical design is key. To extend the analytic value proposition, it will continue to focus on making agile design and implementation easier by offering a Skimmer system with just one S-Blade to iteratively prototype new ideas.

Netezza Customer Case Study: MediaMath

Company Background

MediaMath (www.mediamath.com) is an industry leader in the highly competitive and multibillion-dollar display advertising business. The company provides a demand side platform (DSP) called TerminalOne (www.terminalone.com) that enables ad agencies and large-scale advertisers to identify, bid on, buy, and optimize ad impressions from a variety of sources. TerminalOne analyzes upward of 15 billion ad impressions a day and, using a proprietary algorithm, automatically matches each impression in real time with ads that are meaningful and relevant to users. This involves calculating the fair market value of more than 50,000 impressions every second.

MediaMath is a private company founded in 2007. It has some 110 employees and is headquartered in New York.

For this case study, we interviewed Roland Cozzolino, Chief Technology Officer.

The Business Problem

In a traditional marketing environment, organizations may not see the impact of marketing campaigns for several weeks or months. The BI systems that support and monitor these types of campaigns do not typically have low-latency requirements, i.e., they do not need to collect low-latency data or provide near real-time analytics that enable the business to instantly react to campaigns.

Marketing on the Internet is quite different. The effect of campaigns can be seen in a matter of a few days or even hours. This means that BI systems that support Internet marketing must gather data and generate business performance information much faster than in more traditional marketing channels. The data volumes are also usually much higher. MediaMath's TerminalOne has to analyze more than 15 billion ad impressions per day on behalf of its customers. It needs to monitor and optimize thousands of campaigns that are running at any given time.

MediaMath relies heavily on data to support its operations. Prior to Netezza, MediaMath tried a number of other providers to manage data, but it found that other solutions simply could not keep up with its analytical processing requirements. Queries would sometimes run for days and in some cases would not even finish. The company needed a solution that could support the data volumes and fast responsiveness required in online marketing. It also needed a system that was simple to operate and did not require an army of database administrators to tune and manage.

The Analytic Platform Solution

The system MediaMath chose was the Netezza TwinFin 6. Netezza was chosen after running a POC trial with 5 vendors. Only 2 of the 5 could satisfy MediaMath's performance requirements, and only Netezza offered the simplicity and ease of administration the company was looking for. "The Netezza box takes care of itself, which means our staff can focus on addressing business problems, rather than having to become DBAs," noted Roland Cozzolino. "Success and growth are totally dependent on this solution, and so the fact that Netezza is a public company was also an important factor," he added.

Prior to selecting Netezza, MediaMath also investigated several other approaches, including Hadoop, in-memory databases, and streaming analytics software. These solutions, however, did not provide the required scalability for analyzing hundreds of marketing campaigns. "Given that our revenue has grown by 5 times since December [2009], we needed to be aware not just of current issues but of future scalability needs as well," added Cozzolino. "Hadoop is good for text data and search, but it did not service our needs for manipulating multiple tables of structured data. It also requires the user to have some understanding of the underlying machine architecture. In-memory databases and streaming analytics require too much main memory to meet our performance goals. We need to process and analyze detailed, rather than summarized, data and those solutions simply do not scale well for us." Cozzolino asserts the Netezza FPGA hardware approach is an important factor in achieving good performance.

In-house developed data integration software is used for managing the Netezza data store. The compressed size of the data in this data store occupies between 10 to 15 terabytes of storage. MediaMath achieves about 40% to 50% data compression.

Data in the analytic store tracks details about the ad impressions served to customers, time and geographic information, and how much money was paid to serve the ads. It also tracks the web pages that customers access and products they buy, which provides the ability to relate or attribute purchases to specific ads. This attribution is the key metric used to measure the success of each ad and marketing campaign. All of this data is loaded into the Netezza system in 15-minute batches. This data is kept in the Netezza system for 90 days and is then archived.

Analysis is done using custom-built C++ applications that employ game theory and nonlinear statistics. "Traditional analytic systems, such as SAS, are too slow for handling this amount of data, which is why we had to build our own applications," said Cozzolino. "Both internal users and our clients use the analytics to make rapid decisions, and we need to handle thousands of responses per second, 24 hours a day, and seven days a week. We do very complex processing against huge amounts of data looking for predictive variables and to understand the factors that contribute to successful campaigns. You could hear our old systems crying when trying to do this style of processing." The output of the analytic processing is used to determine the ads that offer the best marketing opportunities and how much to pay for them.

Implementation Considerations

MediaMath found the Netezza platform easy to install, and the integration began producing analytic reports within 3 weeks of the system being delivered. "During installation, Netezza helped us to get things set up correctly, and we quickly scaled from handling thousands of ad impressions to billions without making any significant changes," said Cozzolino.

The current system uses prebuilt predictive models to produce its analytic results, but MediaMath is now beginning to investigate ways of supporting a more ad hoc approach to building analytics in real time to keep it on the leading edge.

To meet future scalability requirements, MediaMath plans to exploit Netezza's ability to push the analytic processing into the database system, which it sees as a key feature for supporting future performance needs. The company also plans to investigate new hardware technologies such as solid state drives.

Benefits

MediaMath is a data-driven company that utilizes analytics to achieve business success and growth. The ability of the Netezza solution to do complex analysis of large amounts of detailed data in a short period of time is an important competitive advantage. This was not possible with previous approaches.

Summary

MediaMath needed an analytic platform that could support fast data loading in parallel with good analytic processing performance. The system MediaMath chose was a Netezza TwinFix 6, which allowed MediaMath to rapidly load and analyze billions of detailed data records. Easy administration and the ability to easily scale the configuration to meet anticipated business growth were also reasons for choosing the Netezza solution. Given the importance of the analytic platform to MediaMath's business, a good working relationship and knowing Netezza is a public and stable company were also critical factors.

ParAccel Overview and Business Description

ParAccel (http://paraccel.com) is one of the more recent entrants in the analytic platform market. After a first shipment in Q4 2007, it has garnered some 30 customers in retail, financial services, pharmaceutical, web media, and other industries as of mid-2010. Like its competitors, ParAccel routinely wins POC engagements against traditional database management systems. It remains small enough that its size is a barrier to adoption for some prospects, but its mid-2009, $22 million funding round is being put to good use adding sales and marketing resources under a new executive team, and its steady growth is testament to its appeal. ParAccel has waved the banner of high performance and used public benchmarks to drive the point home. Among analytic database management system competitors, only Oracle has stepped up to the TPC-H bar. ParAccel released its benchmark under VMware's banner in April 2010, beating out previous benchmarks in the process.

ParAccel likes to say it targets hard problems that demand strong SQL execution and a very effective query optimizer—complex joins such as those that drive retail analyses of market baskets are a good example. It has also worked aggressively to optimize disk usage, partnering early with EMC on an innovative approach that combined retrieval from direct attached storage and a SAN using a patent-pending approach it calls "blended scan." This approach, using the SAN's ability to provide some of the availability, backup, rapid recovery, etc. allows it to offer an enterprise-class data management platform despite its relative newness, but retain its "high performance" model. ParAccel won EMC's Partner Solution 2009 Offering of the Year award. Although EMC's subsequent purchase of Greenplum makes it unlikely that this relationship will significantly drive ParAccel business, the technology was not EMC-dependent and the combined solution's value to customers can easily be replicated using alternative storage platforms.

Architecture

Structurally, ParAccel Analytic Database (PADB) is an MPP system that runs on industry-standard platforms using a columnar-based approach. It may be implemented independently or as part of an appliance-based approach using its Scalable Analytic Appliance (SAA) framework. The latter, when introduced, used "enterprise-class midrange SAN components from EMC," but ParAccel has delivered reference architecture configurations with Dell, HP, IBM, and Sun. It is in the process of incorporating additional storage vendors into its reference architecture.

A leader node coordinates the activities of compute nodes connected via gigabit Ethernet. ParAccel has leveraged Postgres for some of its parsing and planning functions, but is aggressively investing in its optimizer functionality to differentiate itself. Its use of Postgres does not extend to that code line's interpreted query architecture, which some competitors use; PADB uses compiled queries for better performance. Each compute node has direct-attached dedicated storage which can be supplemented with industry-standard SAN- or NAS-based storage. A hot standby node is part of the installation and can step in for any failing node, including the leader node.

PADB's query optimizer boasts several pending patents and is notable for its ability to handle correlated subqueries (CSQs). These feature in several of the TPC-H benchmark queries and have often been a performance stumbling block for less mature products. The engine's ability to remove columns from CSQ plans can have substantial impact on I/O, just as columnar organization does for table scans. Columnar storage aids data compression substantially as well; PADB uses 12 compression algorithms and automatically chooses the correct one. In combination with the benefits realized from not having to use substantial amounts of space for indexing, PADB's ratio of installed storage to raw data is very good and is remarked upon favorably by customers.

"People are asking much more difficult questions; that's why we stress query optimization so we can do many-way joins and CSQs," says CTO Barry Zane. "We don't tell our customers they have to use a star schema. We do better view folding—pruning out irrelevant parts of views and aggregations." The value for its customers lies in the ability to not only ask tough questions, but to conduct an iterative, deductive process that provides true insight at a granular level. PADB stresses its "load and go" nature. Other analytic platforms require ancillary structures to speed up query times. ParAccel's speed eliminates the need for materialized views, elaborate indexing, etc.

Currently ParAccel doesn't provide a management console, instead delivering rich command-line interface (CLI)-based access for monitoring and management. This data can also be displayed using third-party reporting tools from vendors such as IBM Cognos. Management console functionality is planned for future release.

Analytic Functionality

ParAccel's message focuses on allowing users to quickly ask a series of new questions based on the changing needs of the business without months of design, tuning, and implementation. Customers can quickly have a deep "conversation" with their data that provides more accurate insights. Its design offers sophisticated extensibility beyond what SQL language offers for parallel execution. PADB processing of hash joins and aggregates, its architects assert, looks remarkably like the approach taken in MapReduce; the team is complimentary about the way the latter implements functions in the way parallel databases should do them. In general, the stored procedure languages that come with databases can't be parallelized, so ParAccel is pleased that MapReduce has popularized the idea, which it believes plays to its strengths. It expects to offer additional parallelized functions in a future release.

ParAccel touts its SQL performance; it sees CSQs becoming especially common in sophisticated analytics like credit card fraud detection and other applications that look for patterns. Many of its customers are putting detector routines into their systems, running on rules they learned by analyzing the data with ParAccel. At one financial services customer, a single trade over an 11 minute interval relying on highly complex analysis paid for the database.

ParAccel 's emphasis on the power of SQL for descriptive statistics, scoring, and even some correlation analysis has not kept it from enabling such other packages as SAS, SPSS, and KXEN (with whom it shares a board member). It also has customers using R, MatLab, Statistica, and MiniTab. Most of its customers have their own analytic environments in place and want an engine to support them.

Differentiation

Parallelism (along with acceleration) is the core underpinning of "ParAccel." The philosophy of the company is to harness the power of massive parallelism at multiple layers of the analytic stack. Performance is ParAccel's lead story, pointing to PADB's 1 terabyte TPCH record, the first ever virtualized on vSphere, with 7.7 times better price/performance than the prior performance record holder. Its MPP-based platform delivers a scalable approach to performance, which is reinforced by its ability to set the record while using 37% fewer servers. Parallelism also drove a database load time 8.7 times faster than the previous performance record holder.

PADB's reliability is another attribute it highlights; its ability to initiate snaps, which create a point-in-time record of the database for recoverability, allows processing to continue while eliminating the possibility of disaster recovery systems being out of sync. Snaps provide a low-overhead restore in seconds, minimizing the need for time-consuming restores from backup data. Snaps can be scheduled or invoked dynamically; integration with a SAN, managing the snap transaction from within the database, permits support across multiple SAN instances without quiescing the database.

PADB is seeing many varied use cases, including scientific applications such as genomics, proteomics, climatology, and radio astronomy. Customer value drivers for businesses include customer acquisition and retention modeling, lifetime value forecasting, and response modeling. In the financial sector, risk modeling fraud detection, profitability analysis, and simulations are most typical. Other uses are seen in mining and exploration with massive 3-D spatial data problems, in manufacturing with sensor data, supply chain, and RFID applications. Other industries such as telecom, government (especially security) and healthcare are represented in ParAccel's customer base.

Partnerships

ParAccel has established useful partnerships, including a recent agreement with Jaspersoft and Talend to assure 3-way interoperability to create a "data integration plus DBMS and BI" stack. It's also working with IBM Cognos, SAP BusinessObjects, MicroStrategy, and Information Builders on the tools front. Hardware partners include AMD, Dell, EMC, Fujitsu-Siemens, Intel, NetApp, and others.

How Should Customers Start, and What Matters Most?

ParAccel, a fierce advocate of MPP, believes one of the biggest barriers to customer understanding of the new platforms lies with the mind-set of a hardware-based approach. The bottleneck for many analytic workloads isn't the hardware; it's the analytic software sitting on servers and storage. Other vendors paper over performance gaps with massive hardware. Instead, a massively parallel approach using industry-standard servers and storage enables customers to cost-effectively and predictably scale their performance to match future workloads if and when required.

Future/Road Map Exploitation of Trends

ParAccel sees the same trend the market has seen for the past 10 to 15 years—more businesses are defining themselves around how intelligently they use their data. Now these opportunities are available to a wider subset of customers as MPP systems make analytics more powerful, as GUI tools become simpler, and companies like ParAccel offer greater ease of deployment and minimal tuning.

Its own path is toward enhanced SQL extensibility, with scalar, table, and windowed aggregate functions at the top of the list. ParAccel also sees user defined data types (UDTs) being a key opportunity in the longer term. It expects to deliver more mixed workload management in future releases as well.

ParAccel Customer Case Study: Large International Banking Group

Company Background

This large international banking group offers a wide range of services to its over 40 million customers worldwide. The financial application discussed here was developed by a division within the bank responsible for providing financial services to major corporations and institutions. The head of analytics for one of the division's fixed income trading desks was interviewed for the report.

The Business Problem

The trading desk in this case study helps its clients buy and sell asset-backed bonds. The ability to provide fast and informed intelligence about the bonds and their underlying assets is important for traders and salespeople when dealing with clients. This is especially true in today's highly volatile trading environment.

The trading desk saw a significant competitive advantage if they could offer financial intelligence superior to that of their competitors. Many of their competitors, for example, still provide bond and loan performance data using legacy applications that were developed many years before the significant changes that have occurred recently in the finance industry. These applications are slow and inflexible, and cannot easily be adapted to take into account the rapidly changing business climate.

The trading desk, therefore, wanted to develop a new application that could provide superior analytics than those of their competitors. The challenge facing the organization, however, was that its existing Microsoft SQL Server database technology was not capable of supporting the performance requirements of the new application.

The trading desk needed an analytic solution that could handle the ad hoc analysis of billions of rows of detailed data about loans and bonds. The solution had to flexible enough not to constrain users by requiring them to predefine data structures such as aggregates or indexes to achieve good performance. This flexibility was required to enable analysts to quickly introduce new queries to reflect changes in business conditions and trends. This was not possible with their SQL Server system as it took days to test and introduce new queries. The analytic solution also had to handle a large volume of nightly and month-end database updates.

The Analytic Platform Solution

After running a POC trial with 3 vendors, the trading desk chose ParAccel. Key reasons for selecting ParAccel were its ability to meet the project's performance goals and the flexibility it offered in physical design. Particularly attractive were the design of its relational optimizer and the use of a columnar data store.

The other 2 products evaluated were a well-established row-based analytic database system and a less mature database product that offered a columnar data store approach. These 2 products were rejected for several reasons. It was felt, for example, that the row-based system handled scalability issues by "throwing more hardware at the problem." The project group liked ParAccel's focus on software performance, compared with "the brute-force hardware approach" sometimes used by other vendors.

The competing columnar data store product was rejected for its lack of flexibility. The product required the developer to organize the physical layout of the data to suit the proposed queries. The trading desk wanted the flexibility to run different types of ad hoc queries without having to be concerned about the physical organization of the data.

The ParAccel software is installed on a 6-server HP hardware system that supports both active processing and fast backup. The analytic data store managed by ParAccel consists of 2 databases containing approximately 1.5 terabytes of data. Source data for these databases comes primarily from external information providers.

Data loading and analytic processing are done using in-house developed SQL shell scripts. The SQL approach was chosen to provide good performance and to take advantage of ParAccel's SQL support for analytic functions such as lead, lag, and windowing.

The main users of the ParAccel system are three business analysts with good SQL, statistics, and business skills. Their role is to look at business trends, experiment with the data, and produce analytics that traders and salespeople can use to answer trading questions from clients. Salespeople can also submit email requests to the analysts for additional information requested by clients. The analysts then build custom queries to create this information. Any type of information request is allowed. This flexibility gives the 15 traders and 30 salespeople working on the trading desk a significant competitive advantage. Given that the trading desk handles several billion dollars of revenue a year, the cost of producing this information is easily offset by the business benefits obtained.

Implementation Considerations

The organization's IT group was opposed to the purchase and installation of the ParAccel system because it did not conform to corporate IT standards, and this raised compliance concerns. The solution was for the trading desk business unit to take responsibility for the system and for IT to treat it as an experimental black box where "IT just provided the plugs." ParAccel not only provided the database software, but also purchased the hardware on behalf of the business unit. It also provides support for this hardware.

Although ParAccel provided more flexibility than the other solutions evaluated, there was still a learning curve moving from the Microsoft SQL Server environment to the new system. It was important for the business analysts to understand how ParAccel achieves its load and query performance and to code appropriate SQL scripts to exploit the performance gains ParAccel offers. One comment made during our interview was that it was like "giving an Indy car to a Ford driver." The differences between the old and new environments made the running of the POC difficult, and in hindsight, the POC could have been better organized.

The ParAccel system is not as mature as the Sybase operational database system that is implemented throughout the company, but it is ideally suited for sophisticated analytic processing against a large amount of data. There are no database administrators for the project. Instead, the business analysts maintain the system.

The trading desk organization had a good working relationship with ParAccel. Concerns were quickly answered and the support provided by its field engineers was excellent. ParAccel was also found to be very receptive to supporting requests for new features.

Benefits

The ParAccel system provides significant performance gains in both load and query processing times. Month-end loading is now done in about 2 hours. On the Microsoft SQL Server system, the loading had to be broken down into several steps run over several days. The runtime for one of the trading desk's major analytic queries has been reduced to 7 minutes compared with the 3 to 4 days it took in the Microsoft SQL Server environment.

The ability to introduce new queries and achieve good performance is also a major business benefit. Analysts can test new queries and run them in a matter of minutes, rather than having to wait days for the results. It is also possible to run queries that could never be run before. The organization is now delivering analytics to clients that competitors are unable to produce.

The trading desk revenue was significant last year and so the cost of the ParAccel system is small compared to the revenue it helps generate. However, the system had to be justified in terms of achieving the business benefits promised. Now these goals have been met, justifying new hardware to support business growth will not be difficult.

Summary

The trading desk business unit required an analytic platform that could support fast data loading with good analytic processing performance. The business unit chose ParAccel, which enabled business analysts to run existing and new queries to meet customer ad hoc information requirements in a timely manner. This provided the trading desk with a significant advantage over its competitors.
Other important success factors in the project were the ability to generate new exploratory queries without being concerned about the physical layout of the data, easy administration, and a good working relationship with the vendor.

Sybase Overview and Business Description

Sybase (www.sybase.com), acquired in 2010 by SAP, was founded in 1984 and shipped its flagship Sybase database in 1987, competing among the first wave of RDBMS pioneers and differentiating itself with client-server architecture and stored procedures. In 1996, Sybase IQ (purchased from Expressway) was launched as the first modern column store-based analytic database management system. Sybase believes expectations have changed in a business world that is increasingly catching the "Google bug." Business analysts want self-service analysis, typically using unplanned complex queries that must perform rapidly on large data sets. From the outset, this has been the true promise of column stores.

Today, Sybase IQ, available in both symmetric multiprocessing (SMP) and clustered column store database form factors, runs on common platforms such as UNIX, Linux, and Windows. It is installed with more than 1,800 customers; the company likes to point to over 3,200 projects or installations of the product, making it by far the most widely distributed columnar ADBMS. Two hundred new customers were added in 2009. Sybase IQ became available in a cluster version—Sybase IQ Multiplex Grid—in 1999, and Sybase says more than 10% of its installations are clustered (the percentage is much higher in the top 250 sites). Its installed base is well distributed worldwide with all geographies well represented.

Architecture

Sybase believes strongly that for many of its customers, an SMP-based solution is a straightforward choice. Joydeep Das, Director of Analytics Product Management, says, "SMP is simple: one box, any platform. Data placement and distribution, and the management of complex hardware, don't pose any issues." Sybase IQ has been designed to be very intelligent and frugal in its use of hardware resources, including its use of processors and memory. In Sybase's view, the availability of multicore chips has taken away some of the strong arguments in favor of MPP. One socket with 6 to 8 cores in a 4-socket machine—delivering 24 to 32 cores in one box—works well with many reporting and analytics use cases in Sybase IQ's user base.

SMP machines, however, can have practical limitations: you can virtualize, but the capacity tends to be fixed. There are two scale-out choices: MPP or shared disk cluster. Sybase IQ chose not to have to distribute data—all users see the same pool of data in its shared-everything model. Sybase finds that a shared SAN is usually the corporate standard. More users—another department, additional analysts—can be added with another box brought into the cluster. In this architecture, each box supports the new users with the same response time, which differs from the MPP model, where every query tends to saturate every node. Query prioritization and workload management help, but those without priority won't be happy. Sybase IQ in its current offering has not chosen to optimize for the single user with the largest and most complex queries; the company believes that for its customers, single user "light-dimming queries" tend to be an edge case. However, when those complex queries are run concurrently by many users—increasingly a real-world scenario—Sybase IQ's Multiplex Grid cluster architecture almost always wins out.

Sybase IQ has had compressed data since its first release, and this has often led to the perception that it doesn't manage large volumes. By comparison, the company says, MPP and other traditional DBMSs look bloated with secondary indexes, cubes, and materialized views exploding well beyond the raw data. Sybase measures by stored compressed data and says it now has many sites in the 100+ terabyte range even measured that way. Sybase IQ uses Sybase ETL (and other leading ETL tools) and Sybase

Replication Server for moving data in, executing parallel column data inserts, and using microbatching to load trickle feed data efficiently into its column store. Separate read and write nodes facilitate this in the Multiplex Grid Cluster along with pipeline parallelism to help data loads tremendously.

Sybase notes that additions (as distinct from updates) are typical for its customers, for whom load speed has not been an issue. The lack of structure-heavy indexes, cubes, and views also means no updating of those artifacts. Thus, a performance problem for row-oriented systems is avoided, as are the complex administrative tasks associated with it.

Administration via Sybase Central, a mature dashboard with a relatively new user interface, permits cluster management, online database changes, resource monitoring, and security management. Encryption of columns is a key feature demanded by some of Sybase's less than visible government clients (along with FIPS and Kerberos), but has uses as well for its "information service provider" customers who support multiple clients—each can have its own encrypted data. Sybase IQ also provides a powerful web-based monitoring tool called Sybase Control Center that will become the backbone for its next generation administration interface.

Analytic Functionality

Sybase IQ has a library of statistical functions built in and supports SQL-99 OLAP functions and data mining. It enables further in-database analytics via UDF-based libraries and has partnered for including such capabilities as K-means for information clustering, Monte Carlo simulations, and neural networks for classification. Data modeling, reverse engineering of other database designs, and data movement strategies can be modeled in PowerDesigner, a mature and very capable tool. The open source BIRT tool provides simple built-in reporting, charts, crosstabs and other BI functionality. Now part of SAP, Sybase IQ was already well integrated to work with the BusinessObjects family of front-end and data integration tools.

New native drivers for Python, PERL, PHP, and ADO.net extend the product's functionality for web applications. Sybase IQ also supports federated queries, permitting access to information that will never be moved or copied to the data warehouse, an increasingly important use case.

In Release 15.2, Sybase IQ added textual functions, including the ability to search for words and phrases and score the frequency with which a term occurs within a document or text files, permitting analysis of email and other textual information. Leading vendors have largely converged on syntax for textual functions, but today's tools are only the first step in text analysis. While base search and analysis are adequately served by SQL standards—proximity, fuzzy matching, etc.—implementations vary. Sybase notes that text needs to be cleansed and have terms and meanings extracted and so on, and Sybase has the API to plug-in such capabilities from SAP and partners.

Differentiation

Sybase still finds that it competes with "the usual database vendors" for problems where there isn't time or budget to install data into the enterprise data warehouse. For many customers Sybase IQ is used for data marts, but for its information server providers or "data aggregators" who sell data and analytics, such as Experian, Nielsen Media, or Shopzilla, it may be the primary data warehouse on which the business depends. Although smaller vendors are beginning to cross its path, Sybase IQ is one of the few, and by far the most widely installed, high performance column-based analytics server with mature, practical capabilities—the ability to support hundreds of statistical and data mining functions, executed completely in-database, and serving them up through standard application interfaces to concurrent users at high performance levels.

Sybase has also leveraged PowerDesigner to provide a unique information life cycle management capability. With PowerDesigner, customers can specify storage tiers, data retention policies for tables, simulate cost savings with various retention policies and time scales, and automatically generate data partitioning and data movement scripts. The scripts themselves can be managed through Sybase Central.

Partnerships

Sybase's already extensive partner network will be dramatically enhanced by the SAP ecosystem, but already included strong relationships with HP and IBM. The latter helped drive the introduction of the Sybase Analytic Appliance, a series of 6 bundles of software including Sybase IQ, ETL, modeling, loader, backup and recovery with MicroStrategy's BI tool pre-installed on an IBM Power System with IBM Total Storage. Sybase and HP have also recently announced pretested blueprints or reference architectures for two levels (entry and high end) of high performance data warehouse systems on the popular HP DL series of servers running on x86 and Linux. Other tools vendors have partnered with Sybase; SAP BusinessObjects will clearly have an inside track, but MicroStrategy and IBM Cognos have actually enhanced their products to recognize Sybase IQ as a back end and rewrite queries to leverage its capabilities. Fuzzy Logix is the first vendor to be certified as part of the in-database analytics API program; others are expected to follow. IBM DataStage and Informatica have partnered around data integration, and Sybase professional services and systems integrators such as BearingPoint have built Sybase IQ-certified practices.

How Should Customers Start, and What Matters Most?

With a decade and a half of ADBMS experience behind it, Sybase advises its customers not to treat these platforms like their row-based systems. If they approach it with existing skill sets, they will spend time and expense on things they don't need, such as cubes, indexes, and over-provisioned storage. The company recommends that buyers shift skill sets to focus on getting better, faster results that add to business value and reduce overall DBA costs.

Future/Road Map Exploitation of Trends

Sybase sees increasing demand for distributed query processing, self-service provisioning, and increasing interest in cloud-based deployments. It also sees increased opportunities for in-memory databases; some application of those ideas is in the latest version of Sybase Adaptive Server Enterprise, and SAP has made clear statements of its planned directions there. Specifying non-relational problem sets is difficult for the average SQL user, and Sybase is interested in support for other languages for graph traversal, image and text problems often tackled by the emerging NoSQL players. "Why send people to multiple application interfaces to do different things," asks Joydeep Das. "Why not provide extensions to Sybase IQ to deal with non-relational data that can co-exist peacefully with relational data?"

Ultimately, Sybase IQ's goal continues to be in performance leadership with ease of deployment—and its road map reflects just that.

Sybase Customer Case Study: CoreLogic LoanPerformance

Company Background

CoreLogic (www.corelogic.com) is a leading provider of information and services on mortgage financing, servicing, and securitization. Through its LoanPerformance databases, it maintains information on over 100 million active and paid-off mortgages, which are tracked monthly for delinquencies, defaults, prepayments, and foreclosures. The servicing information represents 80% of U.S. residential mortgages, while its securities information covers 95%+ of active non-agency securitized mortgage loans.

The organization's Internet-based services cover risk management, financial analysis, regulatory compliance, and performance prediction. Large financial institutions, mortgage servicers, and lenders use these services for evaluating and predicting mortgage performance and analyzing new market opportunities.

LoanPerformance was founded in 1983 and is now a brand of CoreLogic (NYSE: CLGX). It is headquartered in Santa Ana, California, and has more than 10,000 employees globally. Its 2009 revenue exceeded $1.9 billion.

For this case study, we interviewed Asif Rahman, Director of Application Development at CoreLogic.

The Business Problem

CoreLogic relies on 2 key information databases to support its offerings. The first database is used to manage residential mortgage loan data and is about 1 terabyte in size. The second handles securities-based loans and contains some 200 gigabytes of data. Historically, this data was managed by an in-house application, but was subsequently replaced by a Microsoft SQL Server database system using MicroStrategy analytical processing software.

The challenge the company faced was that as data volumes grew, the existing SQL Server system could not deliver the required performance. IT staff were spending an increasing amount of time building and optimizing aggregate tables and indexes to try and overcome performance issues. Ultimately, the company decided to look for another solution that could meet its performance needs and also support new requirements from its clients.

The Analytic Platform Solution

After running POC trials with several vendors, CoreLogic selected Sybase IQ. Not only did Sybase IQ meet CoreLogic's performance goals, but it also offered the flexibility of supporting commodity hardware. Another key factor in selecting Sybase IQ was its support for the existing MicroStrategy toolset.

"A big benefit of Sybase IQ was that its columnar data store made it easier to add custom data fields to satisfy requests from specific clients," noted Asif Rahman. "Our information databases have grown to in excess of 1,000 data fields and Sybase IQ compresses this extra data, which reduces storage needs for analytical processing and data backups."

In addition to choosing Sybase as the analytic DBMS server, CoreLogic selected Informatica PowerCenter for integrating source data and loading the information databases and MicroStrategy software for creating reports and analytics.

The Sybase IQ, Informatica and MicroStrategy software is installed on 2 HP 16-core SMP servers (4 processors with 4 cores each) with a shared-disk array. Sybase IQ Multiplex is used to give query and update applications running on the 2 servers access to the shared disk array. To comply with corporate policy, the system is configured to provide immediate active-active failover in the event of a hardware or software error. Additionally, access to the two information databases is managed by a combination of MicroStrategy and Sybase IQ security features.

The new system is available 24 hours a day, and at peak usage time about 50 concurrent users from multiple clients access and analyze data in the information databases. Some of these users employ predefined report templates, while others create ad hoc reports. The SQL queries involved are typically highly complex in nature and involve a significant amount of data analysis.

Implementation Considerations

CoreLogic found the Sybase IQ system straightforward to install, and the changeover from Microsoft SQL Server was accomplished easily. The company started with one HP server, and as the customer base and data volumes grew, added a second server and Sybase IQ Mulitplex. CoreLogic also has improved the schema design, dropped aggregate tables, and increased the number of fields in the data stores by a factor of 4. "These changes were very easy to do especially with the latest release of Sybase IQ," noted Rahman. "Sybase IQ was a natural fit for our use case."

At present, most of the analysis is done against numeric data using well-known statistical algorithms. CoreLogic is, however, looking into the use of algorithms from Fuzzy Logix and other companies for more sophisticated predictive analysis.

Benefits

After the initial installation at the end of 2005, CoreLogic saw an 8 times performance improvement and about 40% data compression without making any changes to the analytic applications. Since then the customer base has increased 20 times and the data volume by a factor 2 to 3 times. "To support this growth we only had to add one server," commented Rahman. "Prior to Sybase IQ we couldn't even support 20 customers. The bottom line for us is the ability to meet each customer's query performance needs and grow our customers."

Summary

CoreLogic required an analytic platform that would allow the company to keep pace with its business growth. The company chose Sybase IQ, which provided the ability to scale and support an increasing customer base while also providing good analytic processing performance. Additionally, Sybase IQ's support for the MicroStrategy analytic toolset and an easy migration from Microsoft SQL Server helped contribute to its selection. CoreLogic's Rahman also noted that, "A good working relationship with Sybase was another success factor in the project."

Teradata Overview and Business Description

Teradata (www.teradata.com) is the longest-operating specialist in analytic platforms. Now in its fourth decade, Teradata was founded in 1979 as a proprietary hardware/software solution and was acquired by NCR 12 years later. But 16 years after that, the rising tide of database opportunity made it clear that an independent course was once again a strategic plus, and in early 2007 Teradata was spun out to focus on its position as the leading player in large-scale data warehousing and analytics. It began a steady transition to a new organization and a product line that sustained steady results through the economic downturn until breaking out with two strong quarters at the beginning of 2010 with its best Q2 ever and strong margins and operating income.

Today, Teradata has completed a transition from its one-product legacy to a full portfolio of 5 purpose-built form factors designed to provide analytic solutions at many sizes from laptop to tens of petabytes in the data center and for many workload mixes from small intense usage models to mixed, highly concurrent ones. The product menu now includes a leading-edge, entirely non-proprietary hardware offering, an innovative solid-state appliance, software-only packages, and even cloud-based delivery. The same software, now in release 13.10, runs across all its platforms, facilitating migration from one to another.

The single-node Data Mart appliance, the larger Data Warehouse Appliance, and flagship Active Enterprise Data Warehouse may be thought of as differing primarily in scale, from a maximum of 10.8 terabytes at the low end to tens of petabytes at the top. As customers move up the size spectrum, issues of workload management and administration come into play and the appropriate features are added. Two other form factors represent significantly different design considerations: the Extreme Data Appliance is targeted at very large data volumes (tens of petabytes), but not necessarily the fastest response or maximum concurrency, balancing these against the costs of storage. This model is ideal for small teams working with giant archives and highly complex analytics across very large data sets and is used for tasks like "sessionization" of weblogs, where responding to connected customers in a timely fashion is mission critical. The Extreme Performance Appliance is targeted against what today are more modest volumes—it "only" scales to 17 terabytes, but it uses more expensive solid state disks for the fastest possible response times.

There are multiple software-only editions as well. Teradata Express can run on a laptop; it's free and permits customers and prospects to evaluate, test, and develop at no cost. The Data Mart Edition runs on any symmetric multiprocessing (SMP) machine, and 3 different cloud offerings include an Amazon EC2 version (with a free terabyte) and a VMware version, also free. Existing customers can use the Agile Analytics Cloud product to create and manage "private cloud" sandboxes using available capacity inside existing Teradata environments with minimal IT involvement.

Teradata boasts 1,700 customers in 18 countries worldwide, many of whom have multiple installations and products. Its offerings are used across many industries; the firm markets into 12 major and 23 subcategories with specialized offerings. Teradata's appliance form factor, introduced in 2008, has already garnered hundreds of customers, and the company reports that 42% of these are new names. And although it boasts many of the world's largest forms as its customers, Teradata says that over a third of its customers have annual revenues of less than $1 billion.

Architecture

Teradata may be thought of as an MPP, shared-nothing environment, although it was created before some of today's terminology was in common use, and its descriptions use its own language. Each node (clique) in its system contains multiple virtual engines called Access Module Processors (AMPs), which operate in parallel against its storage environment. Teradata has had decades of experience in managing data transfer from disks mapped to AMPs to ensure that processors are used at the maximum. Processor performance has been increasing faster than disk rates for some time; Teradata's system designs are structured to drive maximum use by mapping storage to processor effectively. (For example, the Data Warehouse Appliance uses 144 drives per cabinet for 4 dual quad core processor nodes.) In the Extreme Performance Appliance this means leveraging the extraordinary I/O of memory-resident data via solid state disks (SSD); millions of rows can be processed in less than a second. Teradata's fourth generation, formerly proprietary-only interconnect, called BYNET, is now implemented on gigabit Ethernet.

Systems are shipped with the chosen operating system (Linux, UNIX, or Windows, with some variations available across the product line) and a bundle of included utilities: loading will automatically drive distribution across the disks, and partitioning, index creation and building, reorganization, and space management are either highly automated or not needed at all. High availability and failover are provided with the Backup Archive and Restore (BAR) offering, a mature, built-in capability. Teradata has added more data compression options in version 13.10, improving its competitive position.

Decades of experience are also evident in the browser-based Teradata Viewpoint management tool, which includes portlets for managing queries, platforms, active systems, workloads, (including dynamic management of workloads), and the database itself: space, nodes, locks, a heat map, etc. Viewpoint users can look at multiple systems in the same view, including ETL processing and multisite operations. Viewpoint is also designed to help manage "data labs" (the sandboxes enabled by the Agile Analytic Cloud offering).

Analytic Functionality

Teradata has a history of innovation in SQL; it was the first vendor to include sophisticated statistical functions within the database engine accessible via its standard SQL interface. The database includes a self-maintaining and managing virtual cube capability to define and process cube-based slice-and-dice type analysis; MDX support for tools is provided via OLAP Connect. UDFs are supported as well.

Teradata Warehouse Miner is another example of early support for "analytics inside." It includes a rich set of analytic capabilities that include data profiling and mining functions, exploration and transformation as well as analytic model development and deployment that execute inside the database without requiring extracts and outside processing, and Teradata is partnering with SAS, SPSS, and KXEN to add additional supported capabilities.

Geospatial Native Database support highlights another leading capability which provides domain-specific functionality, in this case following the SQL/MM standard. Tool partners such as Tableau, ESRI, and CoreLogic have found effective ways to leverage this capability. In addition text analytics, for analyzing, decoding and loading unstructured text data into a relational format to merge with detailed data within the database has long been part of the portfolio.

Differentiation

Teradata's technology clearly helps distinguish it from most of its competitors, but there are other key strategic pieces to the portfolio. Teradata Logical Data Models (LDMs) are industry-specific blueprints for designing an enterprise data warehouse that reflect business priorities. Developed from Teradata's sizable services practice, LDMs organize and structure data to provide a data model and include a subscription program to update the models in future releases.

Teradata's Analytic Applications include data mining/analytics, customer management, master data management, enterprise risk management, finance performance management, demand/supply chain management, and integrated web intelligence. These form a significant piece of the company's business and enhance its strategic value.

Partnerships

Teradata has a large and well-established partner program with a dozens of relationships with system integrators Accenture, Capgemini, Deloitte Consulting, and many others to technology partnerships that often go quite deep, such as its work with Microsoft to develop a "Teradata cartridge" to push SQL back into Teradata for processing and joint work on SQL Server Analysis Services and Reporting Services support. It has been developing in-database analytics with SAS and has a long history of optimizing interfaces with firms such as MicroStrategy, KXEN, Harte-Hanks (Trillium), Informatica, SAP, SAS, TIBCO, and many others. Specific programs exist for advanced visualization partners and data mining partners to develop predictive and descriptive analytics to improve analytical applications.

Future/Road Map Exploitation of Trends

On the heels of its strong results in the first half of 2010, Teradata has laid out a road map of increasing support for in-memory processing alone and with partners, continued enhancements to its geospatial, text processing, and especially its revolutionary support for time as a data dimension. Its work in expanding the use of memory will continue as prices improve, fueling a vision of multitiered storage architectures that will provide automated routing of data to the most cost-effective store, from memory to the least expensive disk hardware, based on policy and usage.

Teradata Customer Case Study: Hoover's Inc.

Company Background

Hoover's Inc. (www.Hoovers.com), a D&B Company, is a research organization that supplies information on public and private companies, industries, and people. Hoover's Online features an information database covering more than 65 million corporations and organizations, and more than 85 million people. The online database includes industry and company briefs, competitive information, corporate financials, executive contact information, current news, and research.

Offering both free and for-pay content, Hoover's focuses on selling subscriptions to marketing, sales, and business development professionals. This content is delivered primarily through Hoover's Online and via co-branding agreements with other online services. The company also publishes its information in reference books.

Hoover's was founded in 1990 and was acquired by Dun & Bradstreet in 2003. It has some 380 employees in North America, Europe and Asia, and is headquartered in Austin, Texas.

For this case study, we interviewed Mamie Jones, Leader Internet Technology, and Bonjet Sandigan, Leader Enterprise Data Warehousing.

The Business Problem

As Hoover's continued to enhance and grow its offerings to add value to the information it offers, the company faced the challenge of not having an integrated view of its customers for sales, marketing, and support processes. Integrated customer information was needed to improve customer retention and to increase revenues by selling new services to existing customers. Integrated customer information was also required to enable Hoover's to analyze different pricing models in what is becoming an increasingly price-sensitive industry.

Customer information was dispersed across many different data stores, and reports were being delivered to users from dozens of systems and applications. Users were frustrated because the results were inconsistent, which made informed decision making difficult. Hoover's decided the solution was to integrate customer data into a data warehouse. The company, however, did not have an existing data warehouse and had little data warehousing expertise.

Hoover's did not have the resources or time to go through a lengthy selection process. Web analytics produced by Omniture were initially considered suitable for producing customer data for the call center, but these were rapidly proving to be inadequate. There was also a risk that business units would try to solve the integration problems themselves, which would lead to a proliferation of inconsistent data marts.

The Analytic Platform Solution

Based on prior experience, Mamie Jones decided that a Teradata 2550 data warehouse appliance would be the best fit to match Hoover's needs. "I was very happy with Teradata in my previous job, and I needed a proven solution that I could rely on and implement quickly," noted Jones. "We needed a data warehousing solution that allowed us to start small, but could be expanded as our requirements grew. The Teradata product set enables us to do that. The availability of Teradata Relationship Manager and the fact we didn't have to do a lot of tools integration were also important for us."

Another factor in deciding to use Teradata was its ability to support 24x7 operations. "We initially need 8 to 5 availability, but we plan to go to 24x7 operations and Teradata's support for high availability was an important factor here," said Jones.

The Teradata platform supports a 3 terabyte data warehouse, but this can be expanded to handle 6 terabytes of data. Data integration is done using software from Informatica. Source data consists primarily of Adobe Omniture web data and Salesforce.com customer data.

Tools from IBM Cognos are used for reporting and analysis. In designing its analytic services, Hoover's separates its users into three categories: weekenders, the curious, and power users. Hoover's completely rebuilt its web portal to support these three types of users. Weekenders primarily use predefined portal objects to create reports. These users also consume prebuilt IBM Cognos reports delivered in Adobe PDF format. The curious group use IBM Cognos to build their own reports and analytic dashboards. Power users use an SQL-level interface to access data and create the analytics they need.

The initial deployment of the system supports a dozen power users and a workload of about a 1,000 sessions per day. Both the portal and the IBM Cognos software have recently been introduced, and this will lead to increased use of the system by less experienced users. When asked about user service level agreements (SLAs) Jones responded, "We haven't published SLAs so far. The users think they are in heaven compared to what they had before."

Implementation Considerations

For a small organization like Hoover's, the deployment of the data warehouse represented a significant investment that involved all the business units working together to fund the project. "Executive management were committed to the project because they saw that a data warehouse was necessary to solve data proliferation issues and to produce analytics for supporting business growth," said Jones.

To get the data warehouse up and running as quickly as possible, Hoover's identified key short-term goals in four areas: better customer acquisition, increased revenue using cross-sell and up-sell opportunities, improved customer retention, and cost reduction. Cost reduction focused on eliminating redundant tools and applications, and on avoiding the building of unnecessary data marts. These cost savings were a significant component of the initial business case for the data warehouse.

"The system was easy to install and get up and running," said Jones. "It is also easy to administer, which means we don't need a full time database administer. The warehouse is managed instead by the developer who supports the ETL software."

When asked what Hoover's would have done differently, Jones noted, "We would not have implemented the Teradata Relationship Manager as aggressively as we tried to do. TRM exposed data quality issues that needed to be fixed in source systems and these issues caused us to miss our launch date for TRM. This was not a technology problem, but a business process issue. For example, if CRM data is entered incorrectly by sales representatives this causes data quality problems in data warehouse."

The upside of the TRM delay was that the data quality issues received the attention of higher management. As a result, Hoover's has now set up a data governance group consisting of both business users and IT staff to address data quality issues. Bonjet Sandigan drives this group with the full backing of the president of the company. "We are still struggling to get users to adopt good data quality practices because there is not a strong data governance culture within the company," said Sandigan. "This was one of the reasons I was brought into the company. We are currently running a data quality education campaign to show people the importance of this. We use the term *data quality* because the concept of *data governance* is seen as threatening and we will get pushback. I think all companies have this problem, regardless of their size."

Benefits

Hoover's deployed the Teradata data warehouse in June of 2009 and achieved a positive return on investment by September 2009. "We continue to monitor the benefits of the data warehouse," said Jones. "Each year during the budgeting process, the business units must predict the revenue they aim to achieve from using the data warehouse. They then have to measure the actual revenue obtained on an ongoing basis."

Summary

Hoover's needed a data warehouse that could provide an integrated store of customer data for sales, marketing, and support. The company chose a Teradata 2550 data warehouse appliance, which provided good scalability and high availability, and enabled long-term data growth.

The Teradata solution allowed Hoover's to quickly deploy a data warehousing environment and a set of analytic services. This prevented business units from building their data mart solutions and incurring additional costs without solving Hoover's data proliferation problems.

Identifying and delivering short-term business benefits to users coupled with removing other ways of creating reports and analytics were key success factors. Data quality issues slowed down the delivery of certain capabilities, but the formation of a data quality management group driven by a data quality expert is beginning to have a positive impact. The support of senior management was also an important success factor for the project.

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 re-run 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 (discussed in this paper) 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 10 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.