InformationWeek
::analytics

Analytics.InformationWeek.com

September 2010
$499

Analytics Report

State of Enterprise Database Technology

Our first *InformationWeek Analytics* State of Database Technology Survey reveals serious fault lines beneath the critically important enterprise database and data warehousing markets, exacerbated by workloads and data volumes that seem to multiply every year, sending costs into the stratosphere. Here's what 755 business technology professionals plan to do about it.

By Richard Winter

Report ID: R1670910

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

Richard Winter
WinterCorp

Richard Winter is an industry expert in large-scale data management technology, architecture and implementation with over 25 years of experience. As founder and president of WinterCorp (www.wintercorp.com), a consulting firm in Cambridge, Mass., he advises user and vendor executives on their strategies and critical projects, focusing on architecture, system engineering and manageability in data warehousing and analytics.

Mr. Winter is an expert in the measurement of database performance and scalability, having directed the development and evaluation of customer and industry benchmarks, pilot programs, and proofs of concept. He has contributed to the development of database products and the implementation of large-scale first-of-a-kind database systems, including systems for real-time analytics and operational business intelligence.

Mr. Winter is a frequent author and speaker, and he teaches seminars on database scalability, including the architecture and selection of data warehouse and data analysis platforms.

Executive Summary

Database management systems are among the most widely used IT products—and among the most profitable for the three software vendors that have for years led this market. In return, IBM, Microsoft and Oracle have provided stability: relational database products, installed on millions of systems worldwide, can claim a coherent mathematical foundation, a nearly universal standard language (SQL) and a large community of professionals skilled in their use.

Thousands of third-party tools and applications employ SQL-based standards to access many different database platforms for virtually every business purpose imaginable.

But just below the surface, the landscape is complex, dynamic and restless. Our *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals reveals discontent among enterprises saddled with rising license and upgrade fees as database sizes and workloads spiral ever larger. CIOs face demanding, and sometimes conflicting, requirements: Manage scale and complexity while minimizing business risk and total costs.

Don't fall into the "data mart of the week" trap because you have no strategy and no capability for integration. Think about security, compression, performance, where NoSQL and extreme analytics fit in, and much more. Oh, and get us that "single version of the truth" yesterday. On the vendor side, Teradata, with sales steadily marching toward $2 billion per year, has a strong presence in the fastest-growing major segment of the database market, data warehousing, an area Netezza has already disrupted with its appliance strategy. Netezza is now a public company growing at 40% per year and challenging larger players in some areas of data warehousing. And, at least 10 other companies, most of them startups and specialty players, are vying for a slice of the database pie.

Executive Summary

New technologies, rapid product development by large and small players alike, and open source products are all affecting dynamics.

In this report, we'll analyze results of our survey and profile the vendors and technologies that are poised to shake up this critical market over the next 12 to 18 months.

Research Synopsis

Survey Name: *InformationWeek Analytics* 2010 State of Database Technology Survey
Survey Date: August 2010
Region: North America
Number of Respondents: 755

Purpose:
To determine the role of database technologies in the enterprise.

Methodology:
InformationWeek Analytics surveyed business technology decision-makers at North American companies. The survey was conducted online, and respondents were recruited via an e-mail invitation containing an embedded link to the survey. The e-mail invitation was sent to qualified *InformationWeek* subscribers.

ABOUT US | ***InformationWeek Analytics'*** experienced analysts arm business technology decision-makers with real-world perspective based on a combination of qualitative and quantitative research, business and technology assessment and planning tools, and technology adoption best practices gleaned from experience.

If you'd like to contact us, write to managing director **Art Wittmann** at *awittmann@techweb.com,* executive editor **Lorna Garey** at *lgarey@techweb.com* and research managing editor **Heather Vallis** at *hvallis@techweb.com.* Find all of our reports at *www.analytics.informationweek.com.*

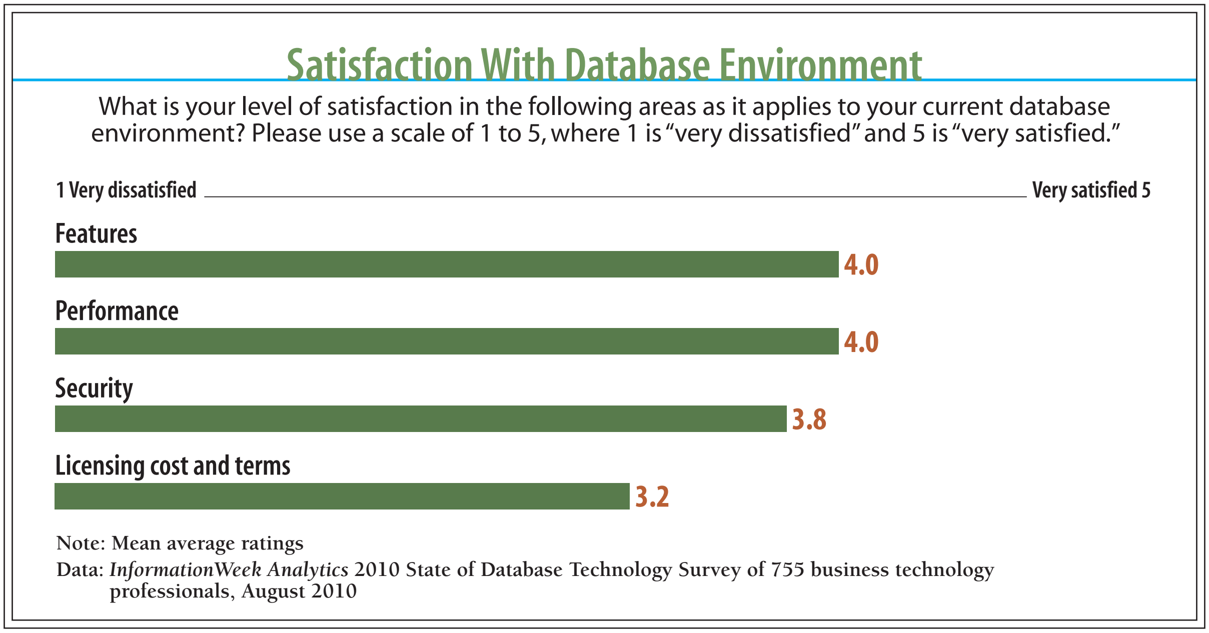
Big Market, High Stakes

We covered a lot of ground in our *InformationWeek Analytics* 2010 State of Database Technology Survey, from operational database management to data warehousing to extreme analytics to security. Some highlights:

- Most respondents, 88%, hail from enterprises where the primary operational database platform is from Microsoft (35%), Oracle (35%) or IBM (18%). While the majority are generally satisfied with features and performance, more than half, 52%, take issue with license fees; 13% of those characterize their costs as "highway robbery." So perhaps it's no coincidence that a remarkably high percentage of respondents, 27%, are using as a secondary operational database the open source MySQL, which is now owned by Oracle and, more importantly, carries no license fee. In addition, 39% are interested in NoSQL, a term encompassing a group of large, clustered but nonrelational data management systems, often inexpensive or open source. Together these trends suggest we'll see movement toward alternatives to the commercially available relational database platforms that have been the near-universal standard for the past 25 years.

- The data warehousing market is also in flux. The good news is that 41% of respondents have a single enterprise data warehouse (EDW) or are working toward that goal—the largest per-

Figure 1



Satisfaction With Database Environment

What is your level of satisfaction in the following areas as it applies to your current database environment? Please use a scale of 1 to 5, where 1 is "very dissatisfied" and 5 is "very satisfied."

Note: Mean average ratings

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

centage pursuing any single strategy in our survey. However, just over 60% are satisfied with the performance and features of their EDW platforms, while a quarter are unhappy with license fees. (Are you sensing a trend?) On the EDW vendor front, IBM, Microsoft and Oracle again lead the list of top suppliers, but other players are much more active in data warehousing, in both EDW and data marts; Teradata accounts for 5% of responses overall and 14% in companies with more than $5 billion in annual revenue, and 11 other data warehouse suppliers show up in our survey responses, signaling the vitality of this market.

- MySQL is frequently cited as a secondary data warehouse or data mart, and in the analytic databases category, which 48% see as distinct from data warehouses and data marts, a remarkable 22% of respondents are using, experimenting with or investigating the open source platform Hadoop; slightly fewer are looking at related tools, such as BigTable or MapReduce.

- Although just 24% of respondents say they're very satisfied with the security of their current database environments, the overall outlook is frankly better than we expected, as 64% of respondents use database encryption, 74% use transaction logging on sensitive data and 70% perform regular security assessments.

Dynamic Space

Database management applications fall into two broad segments: **operational** and **analytical.** Operational applications include transaction processing and some forms of reporting and inquiry, typically against recent data. Analytical applications—also called data warehouses or data marts—involve query, reporting, analysis and data mining, typically against both recent and historical data. In this report we'll explore operational and analytical databases separately. We also discuss a third, emerging segment: extreme analytics, which involves the rapid analysis of extremely large, and often transient, volumes of data.

While the three largest vendors—IBM, Microsoft and Oracle—address all aspects of database management in some way, specialized companies are staking claims within particular segments. Greenplum (recently acquired by EMC), Infobright, Netezza, ParAccel, Sybase, Teradata and Vertica specialize in data warehousing. Hewlett-Packard offers HP NonStop for transaction processing and HP Neoview for data warehousing. InterSystems provides a product, Caché, for high-performance transaction processing. Sybase offers Sybase ASE for transaction processing and a separate product, Sybase IQ, for data warehousing. Meanwhile, the NoSQL movement is a disruptive force to be reckoned with. Some commercial software vendors, such as MarkLogic, which offers a database for unstructured information, have aligned themselves with the NoSQL concept. In addition, there are at least three open source relational database management products on the market backed by commercial vendors: Ingres, MySQL (acquired by Oracle) and PostGres.

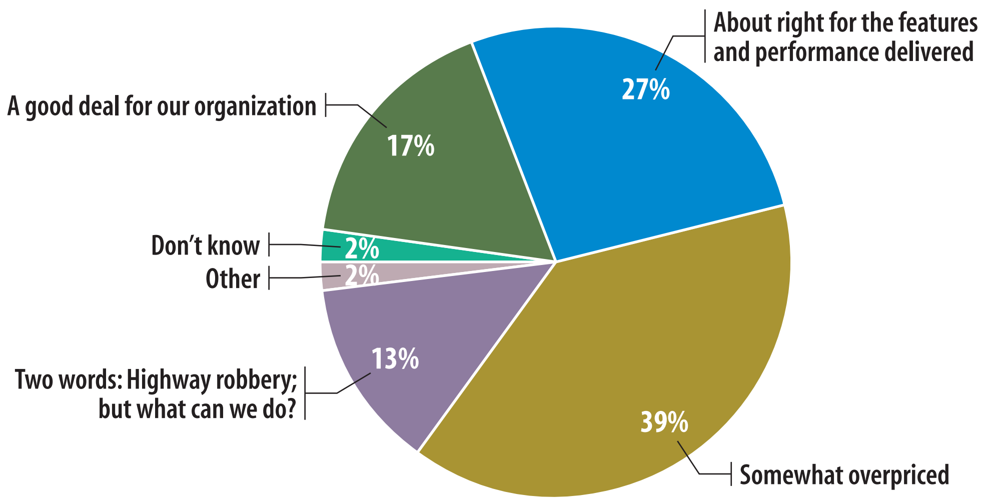
The open source extreme analytics platform Apache Hadoop is in increasingly widespread use. New vendors, such as Cloudera, are offering enhanced distributions of Hadoop, and several others, discussed in more depth below, are addressing Hadoop and/or extreme analytics in some way. Aster Data, Greenplum and Vertica, for example, have introduced ways to support Google MapReduce processing or integrate access to separate MapReduce systems. XtremeData is looking to make a name here, and IBM has an extreme analytics research and services initiative in place as well.

All in all, it's a market in flux. No one ever got fired for buying Oracle, but could an alternate path fulfill business requirements for less? Or will experimenting hurt IT teams looking to delivering fast, unified access to the right data while protecting information?

Figure 2

Licensing Costs for Primary Database

Do you consider licensing costs for your primary database…?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

The DBA's Manifesto

Despite decades spent using this technology, few IT professionals grasp the level of value that comes with really effective database management. This may be partly because of the difficulty of sustaining best practices, especially where support from executives and and business users is lacking. We often see a lack of clarity as to the benefits that may be gained lead to poor decisions concerning database platforms. Our take is that a successful database management program will deliver **15 key elements** of value:

1. Data can be readily located, maintained and retrieved.

2. Data is protected against failures, unauthorized access, theft, vandalism and disasters both natural and manmade.

3. Data from disparate sources can be easily integrated.

4. Data is readily sharable among end users and applications.

5. Known data semantics and data relationships are defined to, and enforced by, the database system rather than in many separate applications.

6. Semantic relationships among data values not recognized in advanced are, once discovered, readily exploited.

7. The meaning and acceptable values of data can be easily and consistently understood across the user community and maintained over time.

8. The timeliness, precision, correctness and consistency of data can be assured according to business requirements.

9. Various users of data can have information presented in the formats, structures and representations that work best for them.

10. Database systems, and the business activities that rely on them, operate correctly, generally perform to service-level objectives, are manageable, and are expandable without unreasonable disruption or expense.

11. A wide range of query types and access patterns are readily supported.

12. Application development is expedited and completed at lower cost because necessary data is readily identified, located and used—and is of predictable quality.

13. Business decisions are more likely to be correct because they can be based on shared, timely facts that are consistent across the enterprise.

14. New, even unforeseen, sources of data may be incorporated into databases and readily integrated with existing data.

15. Application systems run more reliably because they are not disrupted with unavailable, incorrect or unexpected data.

Together, these statements define the ideal result of a long-term database management program. Many require a commitment of human expertise in addition to system capability, of course, but the choice of platform dramatically affects the likelihood you'll achieve these outcomes. For example, in data warehouses, guard against placing too much emphasis on the platform cost—especially the upfront software license fee—and too little on the product's functionality. If your platform doesn't support integration well, and doesn't make it easy to incorporate new data sources as they show up in business requirements, you will end up buying a second platform for the new data that shows up a year after you build your data warehouse, and that beat could go on. This is one reason some companies have hundreds of data marts, most maintaining unconnected silos of information.

Marts vs. Warehouses

Individual data marts are fine—when that's what the business really needs. But create data marts to support occasional special situations, not as your principal way of operating or because your data warehouse platform has failed you. A modest number of data marts is a sign of a vibrant environment. Many data marts, or a high rate of data mart creation, is a disaster waiting to happen. You'll face widely replicated data and inflated system costs that can run into the tens or hundreds of millions of dollars in a large enterprise. And the cost in lost business opportunity of fragmented data is typically 10 to 100 times higher than even the inflated system costs of data replication, possibly amounting to billions of dollars in the enterprise. Clearly, data integration should be Job 1. So, what do you need to support integrated data across a set of always growing subjects and uses? Isn't that what data warehouse products are supposed do? Well, they all do it in some way, but if you have robust data integration requirements, then you need to look for:

- Good performance on complex queries that join data from multiple large tables distributed on different keys; this means good performance on noncolocated joins in which there's more than one large set to be joined.

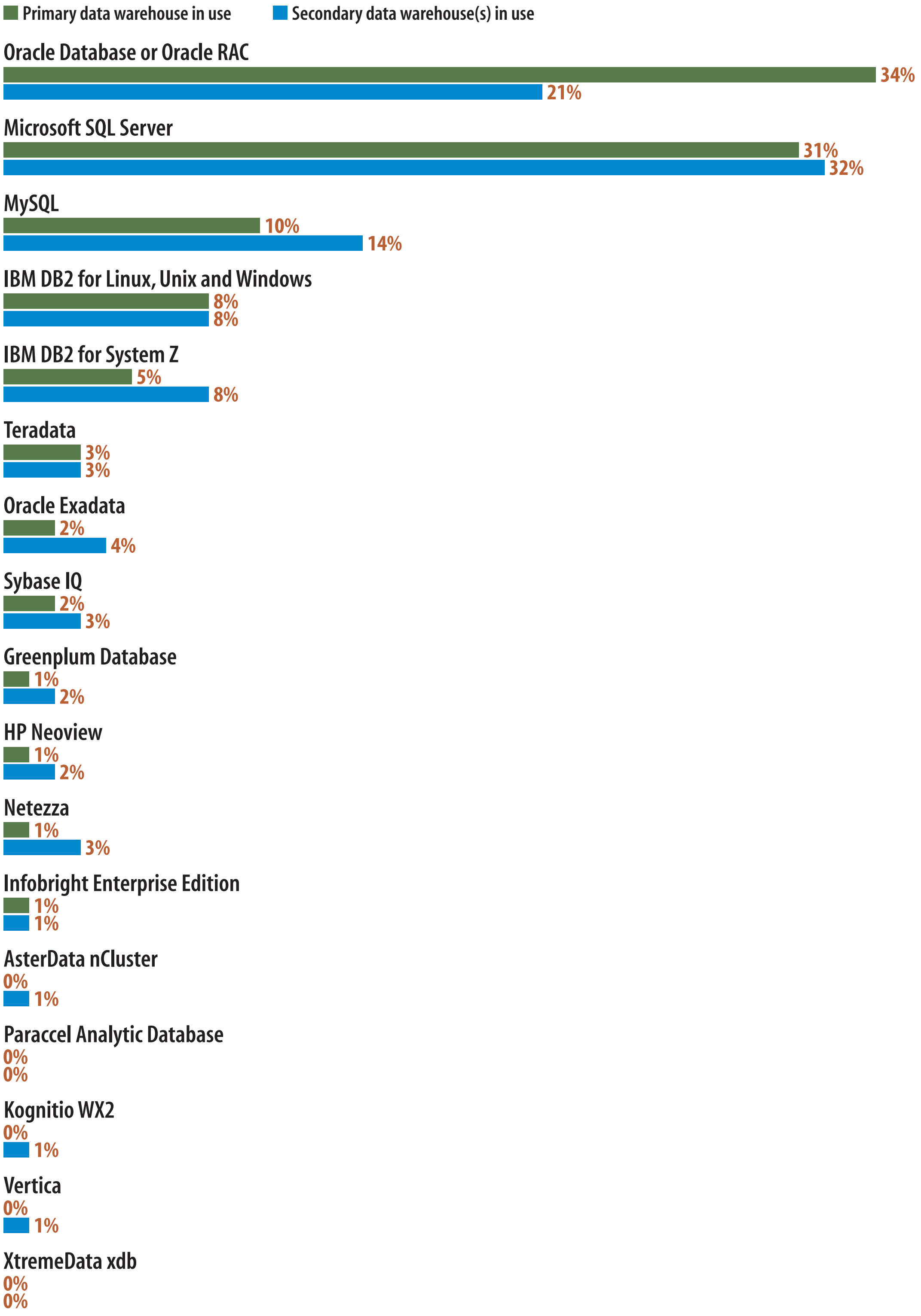
- Strong support for an evolving schema to which you can readily add lots of views, tables and relationships, while also changing virtually any aspect of existing tables, views, columns and relationships with little or no interruption of ongoing database operations.

- The ability to maintain performance aids transparent to the user (think materialized views, indexes, clustering) and transparently change them as database usage patterns change, also without interruption.

Figure 3

Enterprise Data Marts Currently in Use

Which of the following is the primary enterprise data mart currently in use at your organization? Which are secondary in use?



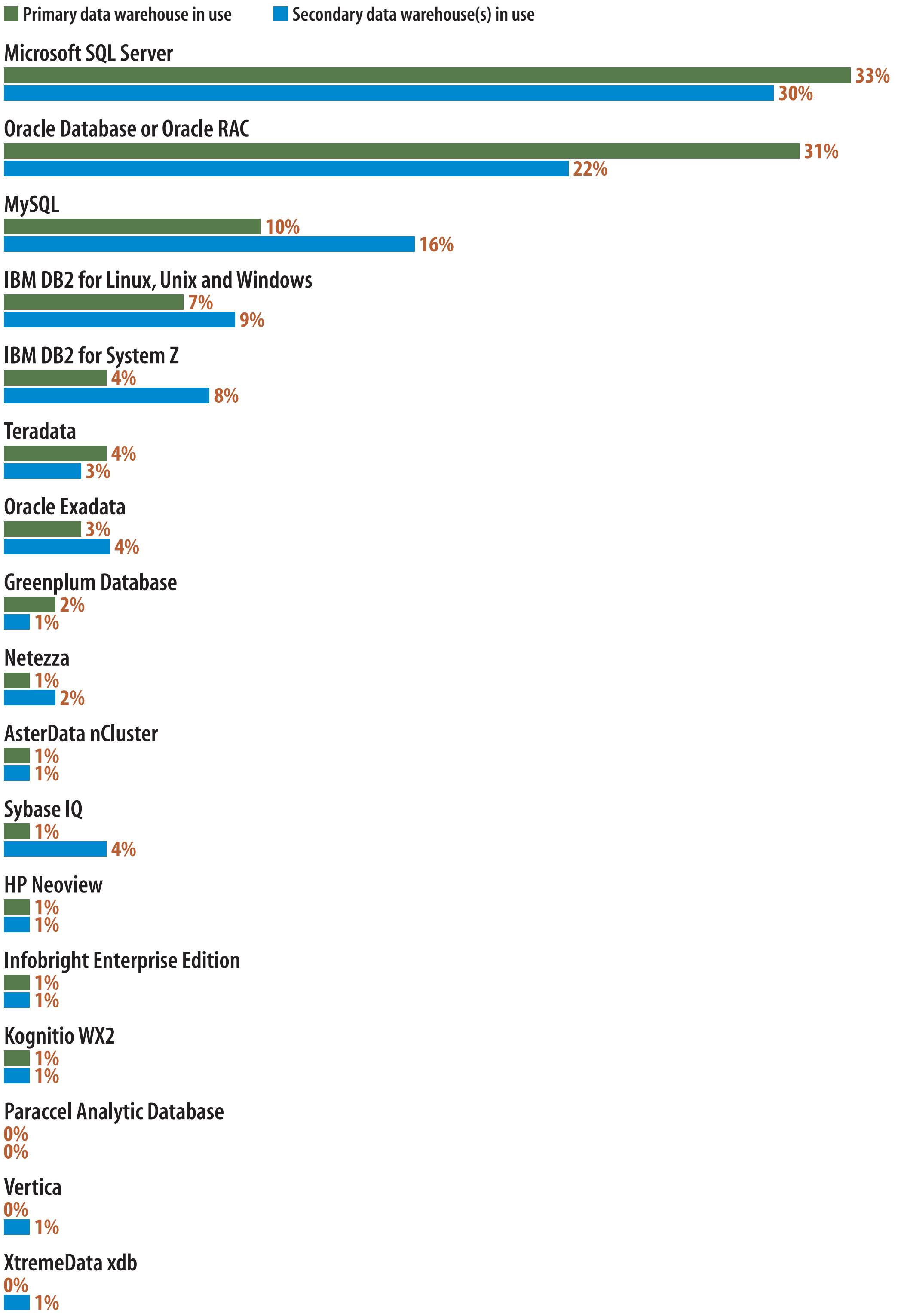
Base: 422 respondents at organizations with one or more enterprise data marts

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Figure 4

Future Use of Enterprise Data Marts

Within the next 12 to 18 months, which of the following do you expect to be the primary enterprise data mart in use at your organization? Which do you predict will be secondary in use?



Base: 422 respondents at organizations with one or more enterprise data marts

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

- A smart optimizer that can handle all this complexity and change and still, most of the time, get queries to perform efficiently on its own.

Yes, it's a tall order. But the goal of integrated data shared across a substantial, dynamic set of uses is a worthy one. In contrast, the requirements for data mart platforms are less stringent than for an enterprise-class data warehouse. Many can use a star schema and be built around a single large fact table. The queries and workload mixes are typically less complicated, and data availability requirements are often lower. Data marts are a good place to try out new products and technologies that may fit certain applications.

Bottom line, data marts created to meet a pressing business objective are fine—as long as they don't derail an ongoing program to support integrated data.

Changing Face of Database Technology

Though anchored by the standard SQL language, which changes only once every few years, database management products are the focus of intensive R&D programs. Major cost- and labor-saving advances we've seen in recent years include:

- The appliance model, which increases the speed with which new systems can be implemented and reduces costs.

- Broader acceptance of parallel architectures, particularly shared-nothing massively parallel processing (MPP), which increase scalability, performance and modularity.

- Data compression schemes that exploit rising processor power to store, read and write data more efficiently.

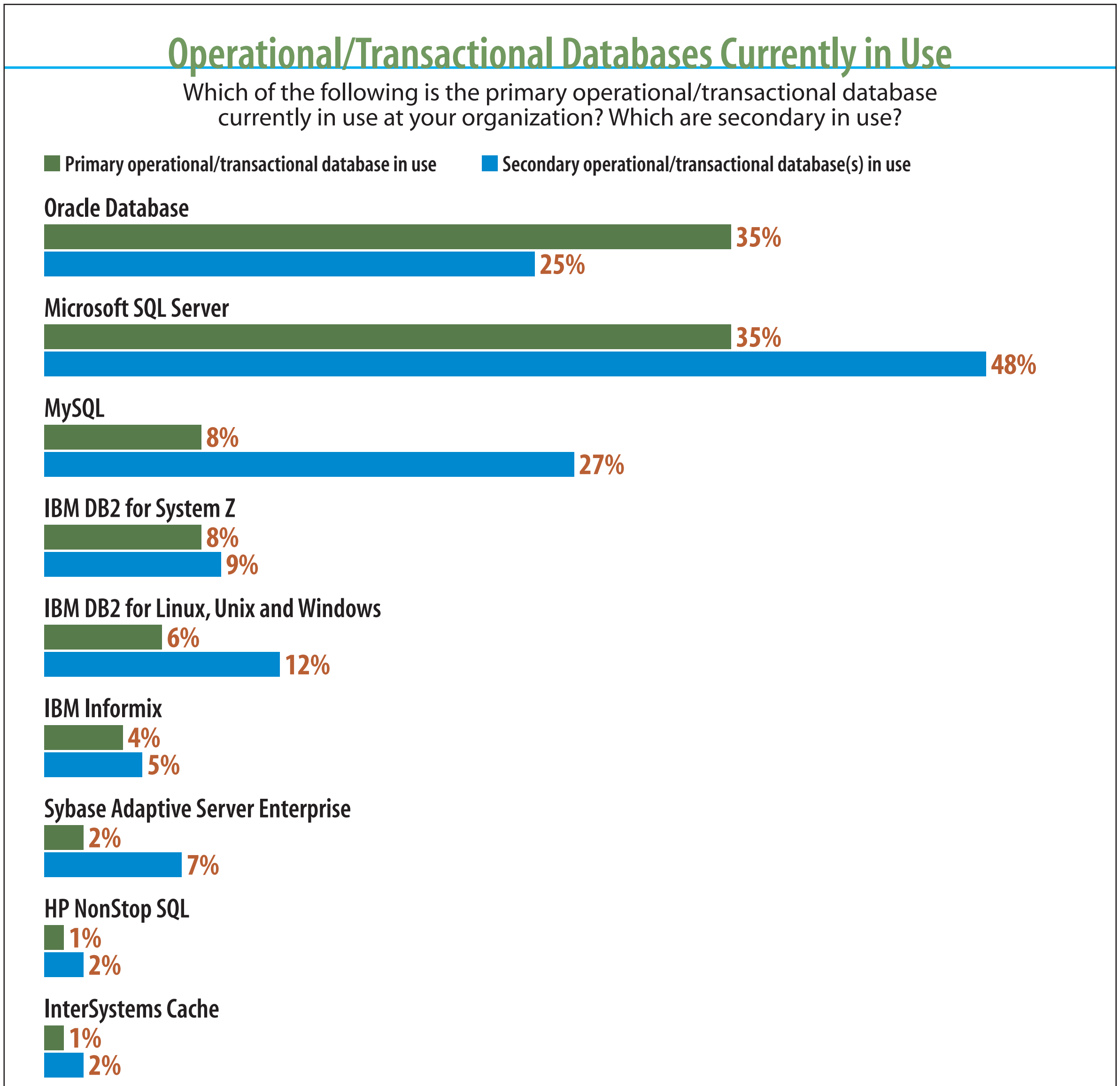
- Solid state or flash disk, which increase storage bandwidth and lower latency to provide higher performance.

- Sophisticated data partitioning and clustering methods, which reduce work and increase manageability.

- Systems management automation, particularly in such areas as mixed workloads, performance management and troubleshooting, to slash maintenance costs.

Among these, we expect that flash storage, also referred to as solid state disk, will have the largest impact. While SSDs have been used in specialized appliances and high-end PCs for awhile, the past 24 months has brought a significant rise in enterprise storage systems employ-

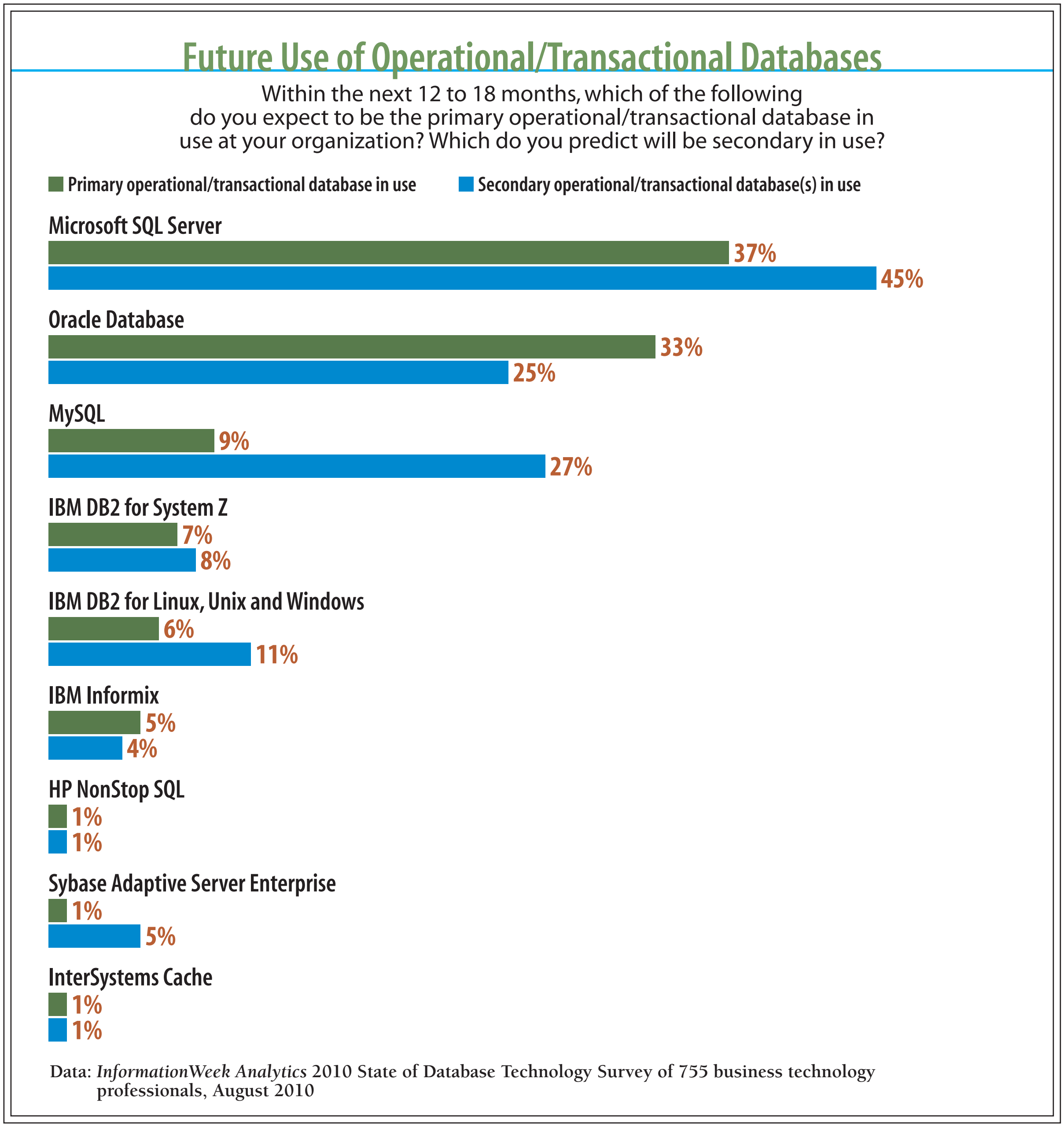
Figure 5



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

ing this technology. In general, SSD performance is 10 to 100 times faster than with spinning disk, depending on the application, and space and power requirements are lower. The downside: Cost is also about 10 times higher per GB of storage than with spinning disk.

Figure 6



Future Use of Operational/Transactional Databases

Within the next 12 to 18 months, which of the following do you expect to be the primary operational/transactional database in use at your organization? Which do you predict will be secondary in use?

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Since last year, database vendors have been delivering products that incorporate flash storage, usually in hybrid configurations that also use spinning disk. Typical ratios are about 2% to 15% flash storage. Even this modest amount can deliver more I/O capacity (reads or writes per second) than the much larger complement of spinning disk drives with which SSDs are packaged. That's because spinning disk drives perform, at most, a few hundred operations per second, while flash drives are capable of tens of thousands per second. Variations on this approach are employed in Oracle's Exadata 2 and in IBM's Smart Analytic System. For high-performance analytic applications, Teradata has introduced its Extreme Performance Appliance 4600, a data warehouse system that uses entirely solid state storage.

We expect to see improvements in both performance and price competitiveness over the next few years. Eventually, SSDs will be the principal medium for high-performance data storage.

Advances notwithstanding, our survey respondents are clearly chafing at licensing costs. While 79% are either satisfied or very satisfied with the features and performance of their operational database platforms, only 39% feel the same way about licensing terms. More than half consider their primary databases overpriced. And, bigger but less visible financial factors, such as the hidden cost of fragmented, duplicative and unmanaged data, add pressure.

Cost is not the only factor influencing the dynamics of database practice, either; respondents cite spectacular growth in data volumes and an increase in sources that must be supported. Escalating requirements for rapid, near-real-time decision-making and continuous operation "no matter what" are causing many to rethink architecture and platform choices to increase data availability and reduce latency. In systems that are sufficiently large, complex and critical, scalability, manageability and security may trump other factors. Let's dig into key database areas and discuss decision points.

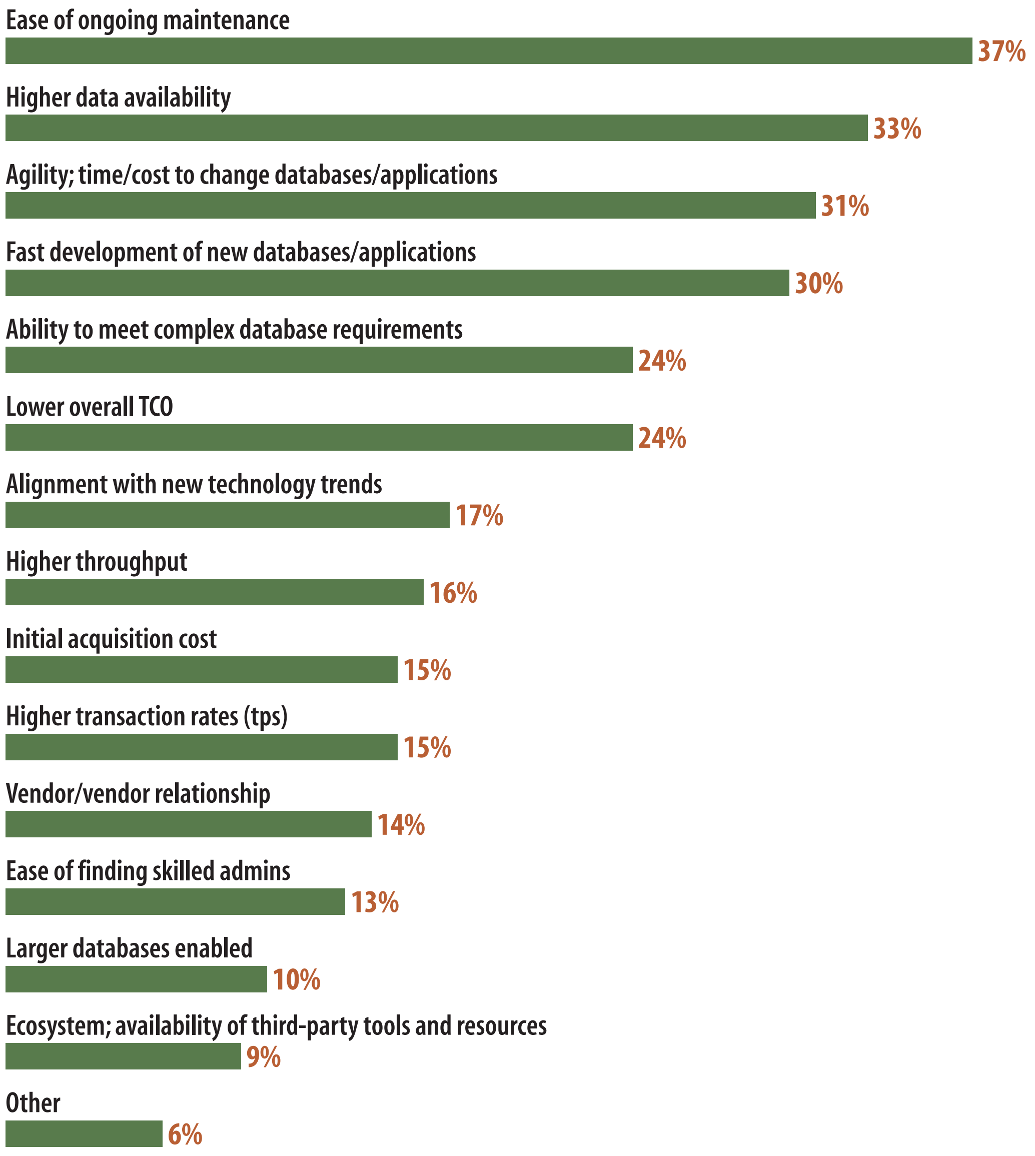
Getting the Job Done

Operational databases support the transaction processing of day-to-day business matters in virtually all enterprises: processing and filling orders, recording shipments received and sent, doing the accounting and payroll, and controlling production. The director of operational databases at a large high-tech company once introduced his team as, "The folks who make sure the laundry gets done." Point being, no business can keep going for long unless someone makes sure the operational databases are online, intact and performing.

Figure 7

Factors Influencing Choice of Operational Database

What are the top factors that influence your choice of operational database?



Note: Three responses allowed

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Survey respondents say four factors most frequently influence their choice of operational database platform. Three of these four relate to the ease or speed with which databases can be implemented, changed or maintained. In a word, this is about agility: rapid response to a changing world. At No. 4 is data availability—having ready access to the data needed for transaction processing, 24/7/365.

Key factors in platform selection: The three vendors with more than a $1 billion each in database license revenue—IBM, Microsoft and Oracle—account for more than 80% of our respondents' primary and secondary operational databases, and we see no indication data that their presence will decline significantly over the next 18 months. Two interesting data points here: Microsoft SQL Server is by far the most frequently cited secondary operational database, with 48% of respondents identifying it in this role. And, MySQL is cited as the primary operational database by 8% of respondents and as a secondary operational database by 27%. (Respondents were allowed to indicate multiple secondary operational databases in use.) Thus, survey respondents cite MySQL as a secondary operational database more frequently than any other except Microsoft SQL Server.

Among respondents from organizations with more than $5 billion in annual revenue, the percentage using MySQL as a secondary database is not quite as large, at 18%. But, that percentage rises to 22% for those big companies expecting to deploy the database within the next 18 months—a significant presence. IBM platforms, both on the mainframe and on Linux, Unix and Windows, also play a large role among respondents with revenue over $5 billion and demanding requirements for operational databases, including large workloads, high transaction rates and complex databases, as do Sybase ASE and HP NonStop SQL.

Still, it's clear that use of open source database management systems is growing at orgs large and small, although typically not on the largest scale or most critical applications. Though MySQL got a lot of attention in our survey, there are two other open source relational database systems in widespread use: Ingres and PostgresSQL.

Ingres is one of the earliest relational databases and was on the market for more than two decades as a commercial product. Within the last few years it has become available as open source from Ingres Corp. and is now offered alongside Ingres VectorWise, a new product for analytical database applications.

PostgresSQL is an object-relational open source database system first offered as an open source product around 1995. It's been incorporated in some form in several data warehouse products now commercially available and is in widespread use, with robust sources for support and hosting services.

Interestingly, Ingres and PostgresSQL both have roots in the University of California at Berkeley, where Professor Michael Stonebraker and his graduate students initiated these database projects along with others in the 1970s and 1980s. Stonebraker also founded Vertica and VoltDB, two more-recent commercial products we'll discuss later.

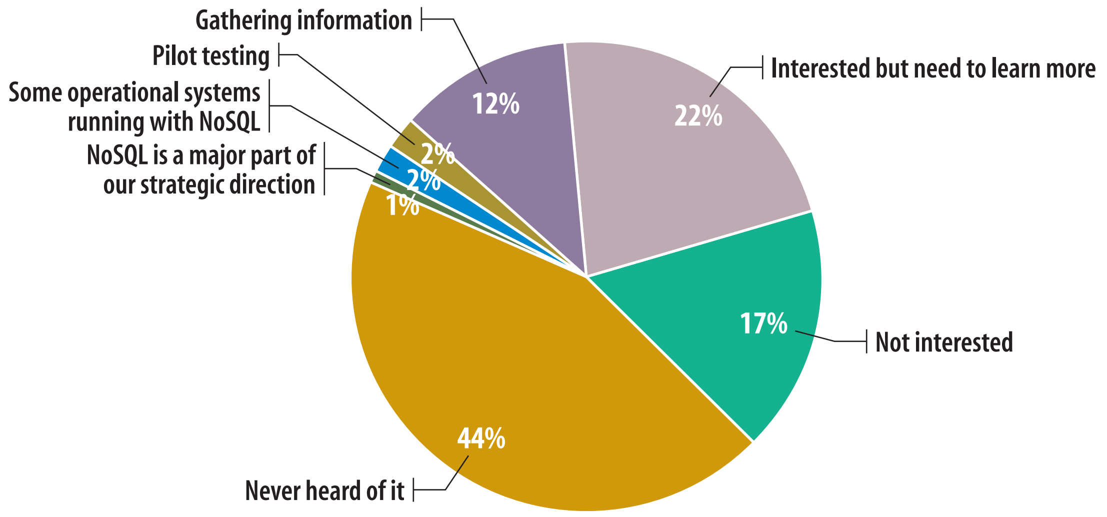
New Alternatives From the NoSQL Camp

The term "NoSQL" in its present usage gained popularity in 2009 and is now associated with a wide range of open source and commercial database products that in some way differ from the relational database systems that dominate the market. The Web site http://nosql-database.org/ lists over 70 data management systems that it says fall into the NoSQL category. Some of these

Figure 8

Interest in NoSQL

What best describes your degree of interest in NoSQL?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

are commercial products that existed before “NoSQL” became a term, such as EMC Documentum xDB, MarkLogic, Objectivity and Versant. But others, such as Cassandra, are open source offerings that emerged from more recent efforts by large user organizations.

With respect to operational databases, the term “NoSQL” applies principally to data stores that:

- Are not applying the same rules of data consistency, that is, ACID rules, as are commonly employed in relational databases doing transaction processing;

- Are not employing SQL as the query language; and

- Are storing something other than structured tables, typically documents, graph structures such as those arising in social networking, or loosely structured records defined as sets of key-value pairs (and hence able to vary in structure from record to record and over time).

Those involved in NoSQL efforts face daunting new requirements for operational database management, an area they felt was not well addressed by the relational database systems available. Much of the development and initial use of this technology has occurred in large Internet-based businesses, such as Amazon, Facebook, Google and Yahoo. Cassandra, now available free from the Apache Open Source Foundation (cassandra.apache.org), is a leading example of a NoSQL platform suitable for large-scale operational systems. It is in use on a distributed architecture across many servers at Digg, Facebook and other sites.

As shown in the figure, previous page, 39% of our respondents are interested in NoSQL. Approximately 5% are using some data management platform associated with NoSQL, and 12% are gathering information to determine its relevance to their data management directions.

Another new answer to high-performance transaction processing requirements, but one that does employ relational database technology and enforces ACID rules, is VoltDB. A startup company founded by UC Berkeley’s Stonebraker, VoltDB aims to satisfy some of the same requirements addressed by NoSQL, but with a different approach. VoltDB is open source, though supported by a venture-funded startup. It sports a distributed architecture and supports demanding requirements for performance and scalability in OLTP. Unlike the NoSQL platforms, VoltDB promises to deliver high performance while continuing to manage data consistency, relieving the application programmer of that burden.

Data Warehousing: Foundation for BI and Analytics

A data warehouse is a database that exists strictly to support decision-making, query, reporting and analysis—a set of uses often bundled as business intelligence.

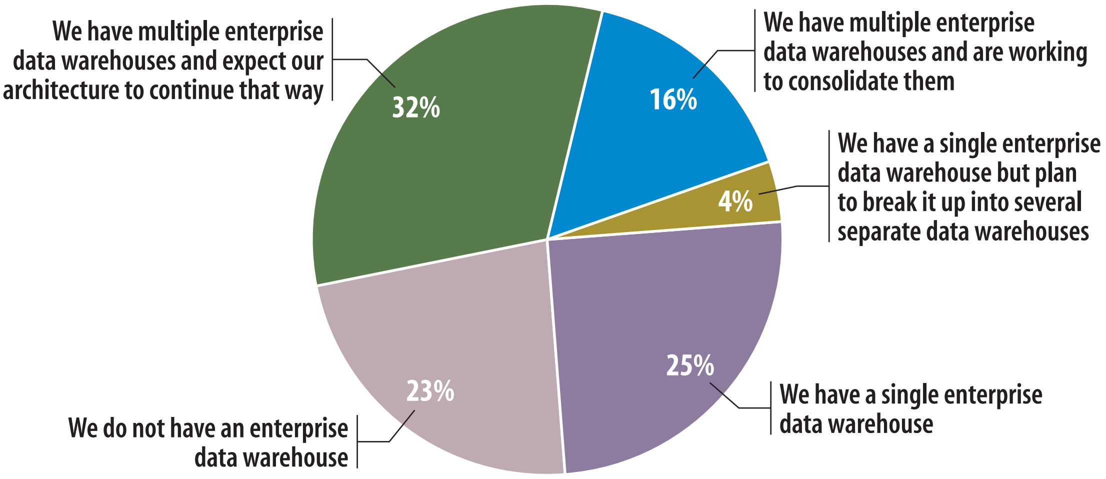
Data warehouses typically store a copy of all data originally created in operational databases. For example, an operational database may make a duplicate of each customer order as it is created or received and hold this data until the order has been fulfilled. Depending on the business, that might take anywhere from a few minutes to a few months. However, most businesses keep a long-term history of orders in a data warehouse, retaining them for a period of years for reporting, analysis and perhaps compliance purposes.

In addition to operational data, data warehouses often also store other information obtained from diverse sources. For example, a company that markets products or services to consumers might purchase data about people likely to have an interest in the company's offerings and

Figure 9

State of Enterprise Data Warehouse

What is your enterprise data warehouse situation?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

store it in a data warehouse, enabling business analysis of which consumers are most likely to be interested in a specific offer or new product.

Data warehouses, and the products used to manage them, vary in scope. As discussed previously, a data warehouse created for one specific purpose and holding just one narrowly defined class of data is properly called a data mart. The term "data warehouse" is reserved for systems that store data on multiple subjects and support multiple uses. Thus, a data warehouse might be created to support all financial analysis and decision-making. Such a warehouse would contain all cost and revenue data in the organization and could contain data on several other subjects employed in financial analysis, such as products, stores and suppliers.

An enterprise data warehouse (EDW) provides a single, integrated repository to manage all the data used for decision-making, query, reporting and analysis. By having one central repository, where such data can be brought together, cleansed, quality-controlled and integrated, an enterprise can realize large savings. Duplication is reduced. Quality is increased. All users of the data get the same answer to a given question, the proverbial "single version of the truth." Data can be cleansed once and used all over the enterprise.

While the advantages of a well-run EDW are formidable, not every company will find it practical to create one. As a result, a variety of approaches to data warehousing are in use. Forty-one percent of our respondents work in enterprises where either they have a single EDW or are working to create one, as shown in the figure, previous page. About another third have multiple data warehouses that are regarded as having enterprise status. Sometimes, a decentralized organization or an enterprise consisting of several highly autonomous business units will maintain multiple, separate "enterprise" data warehouses. The remaining respondents work in organizations in which there is no enterprise data warehouse or direction toward one.

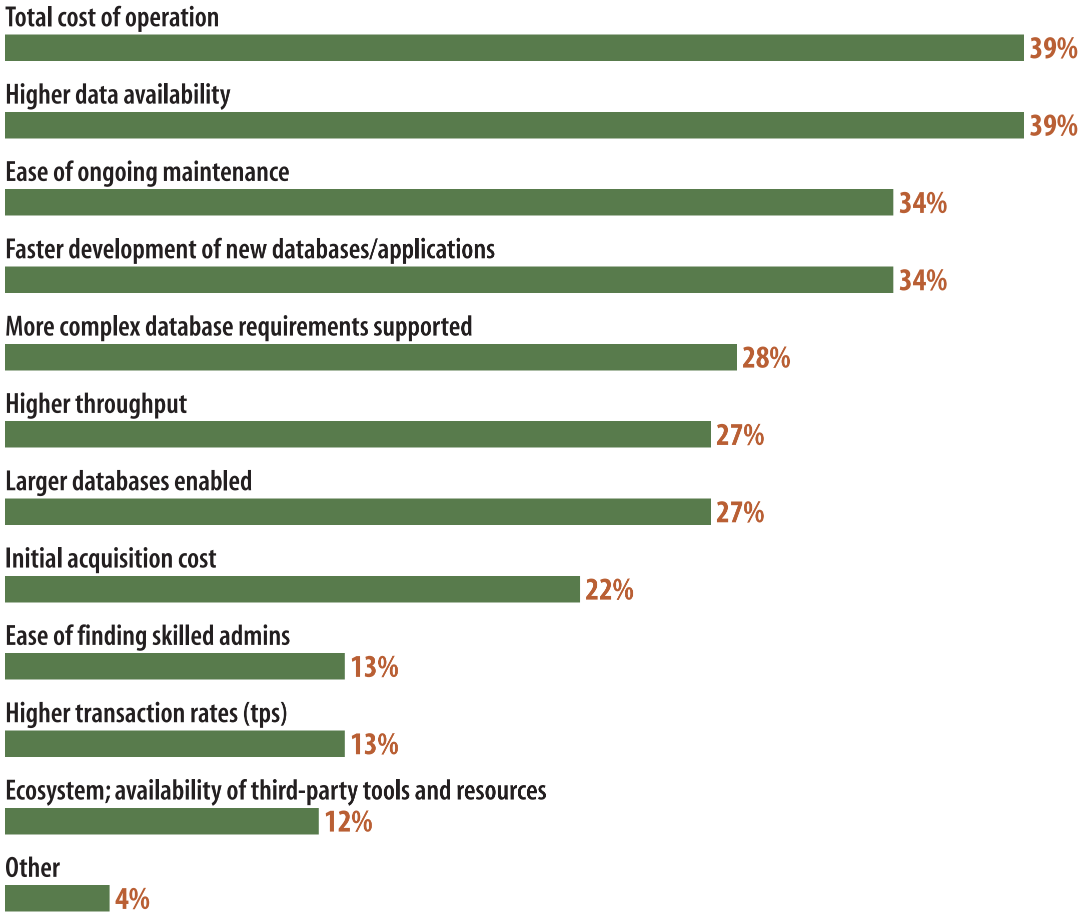
When selecting a platform for an enterprise data warehouse, the top two concerns of our survey respondents are total cost of operation and data availability. When selecting a platform for a data mart, the top two concerns are fast development and agility. This makes sense, as data marts are typically created to accomplish a specific business objective, often under time pressure. By contrast, an EDW is usually a larger undertaking aimed at economically meeting a range of business needs over a longer period of time. If an EDW supports a substantial community of users, the totality of their requirements is likely to mean that there is rarely, or perhaps never, a good time for the data to be unavailable.

Thus, it is increasingly common for EDWs to require high or continuous data availability. Furthermore, because an enterprise EDW will often grow to manage data on many subjects and support a substantial workload across different business units, it is often subject to more scrutiny with respect to cost.

Figure 10

Factors Influencing Choice of Data Warehouse

What are the top factors that influence your choice of data warehouse?



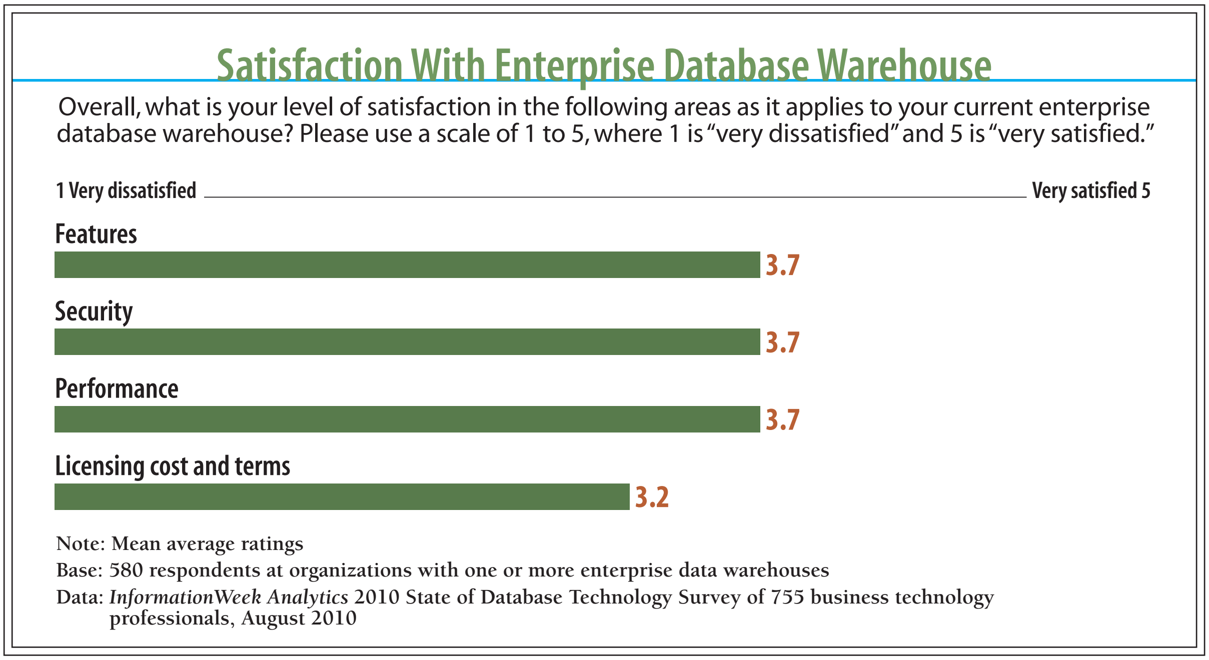
Note: Three responses allowed
Base: 580 respondents at organizations with one or more enterprise data warehouses
Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

There are a number of other requirements typical to the EDW that are not present for everyday data marts:

- An EDW is usually called on to integrate data across many subjects, meaning that it must support a complex schema. Further, the range of subjects and uses tends to expand over time. Often, the best approach for this situation is a normalized schema, where relationships among entities are represented neutrally, making it easier to represent a variety of data relationships and queries. For example, in a database about patients, doctors, diseases and treatments, a normalized schema would treat each of these as an independent entity, not defined in terms of the others. Queries that focus on doctors are therefore no harder or easier than those that focus on patients or diseases.

- An EDW requires the ability to support such a normalized schema and also often requires a physical design in which there are multiple large tables that cannot be distributed on a common key. Both these design considerations are consequences of the role of the EDW: supporting all analytical uses of the data across the enterprise. Data marts, because they are more purpose-specific, can frequently employ simpler database designs.

Figure 11

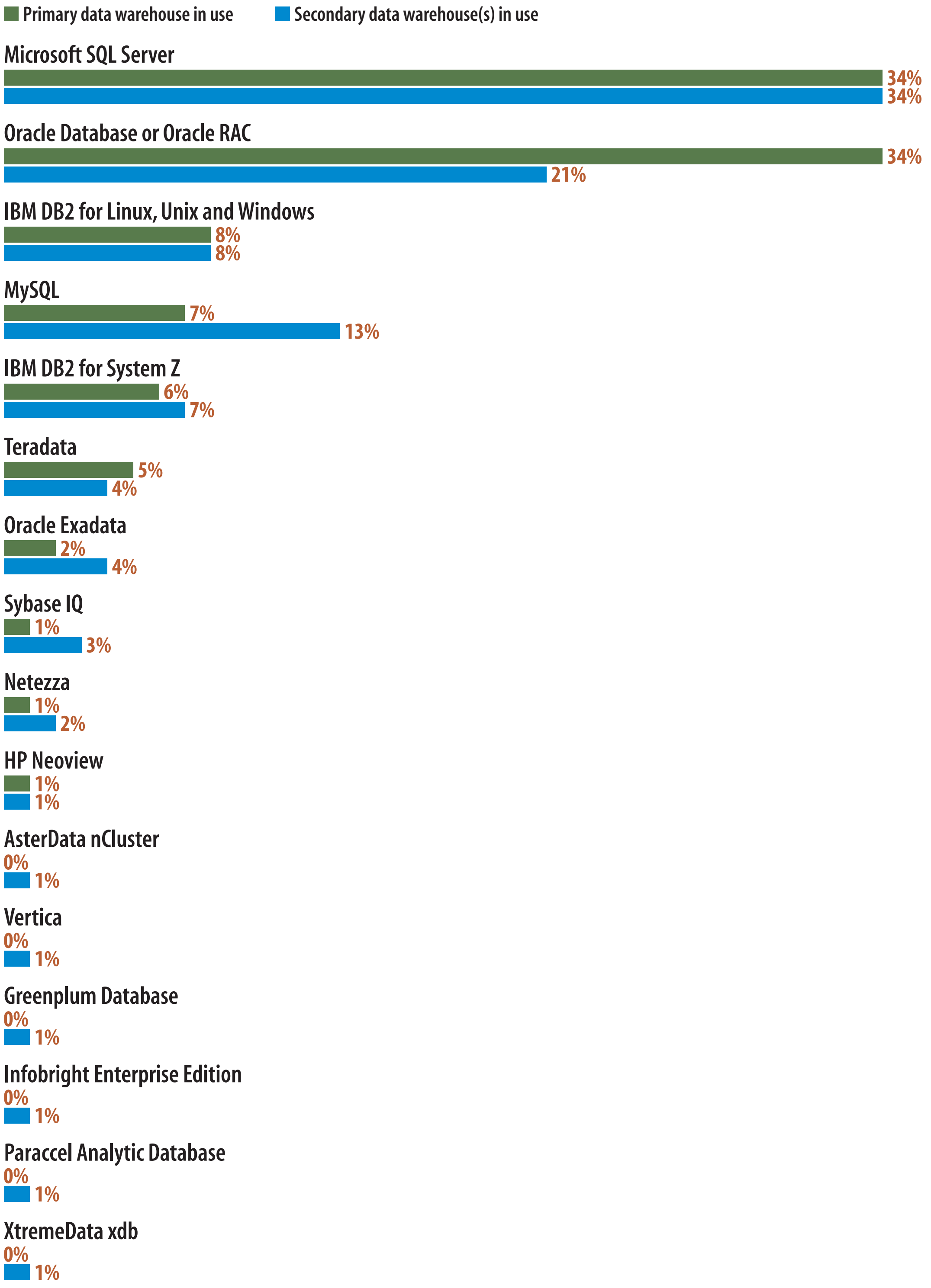


Satisfaction With Enterprise Database Warehouse

Overall, what is your level of satisfaction in the following areas as it applies to your current enterprise database warehouse? Please use a scale of 1 to 5, where 1 is "very dissatisfied" and 5 is "very satisfied."

Note: Mean average ratings
Base: 580 respondents at organizations with one or more enterprise data warehouses
Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Figure 12



Enterprise Data Warehouses Currently in Use

Which of the following is the primary enterprise data warehouse currently in use at your organization? Which are secondary in use?

Base: 580 respondents at organizations with one or more enterprise data warehouses

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

- In addition, the workload for an EDW is often a mix of complex and simple queries, short and long-running queries, and both large batch updates and frequent online updates. The broad requirements call for comprehensive capabilities, the large and varied workload puts a premium on system stability, and the complexities require a highly capable optimizer.

Often, the entire business relies on the EDW, resulting in unforgiving expectations and a risk-averse approach to selecting the platform. This comes through in our survey: 61% of respondents are somewhat or very satisfied with EDW platform performance, and 63% are similarly satisfied with EDW features, compared with 75% and 79%, respectively, for operational databases. Similarly, only 37% are somewhat satisfied or very satisfied with EDW license cost and terms. Thus, EDW users, facing demanding and complex requirements, are still looking for more capability at a lower price.

While the leading operational database vendors—IBM, Microsoft and Oracle—also provide the EDWs for many respondents, the overall product lineup is quite varied. In particular, there are a substantial number of systems designed specifically for data warehousing. The most widely used is Teradata, the primary EDW product at 5% of our respondents overall and at 14% of respondents from companies with revenues of over $5 billion annually. IBM DB2 for Linux, Unix and Windows; Microsoft SQL Server; and IBM DB2 for System Z are also popular EDW platforms.

Beyond Toasters

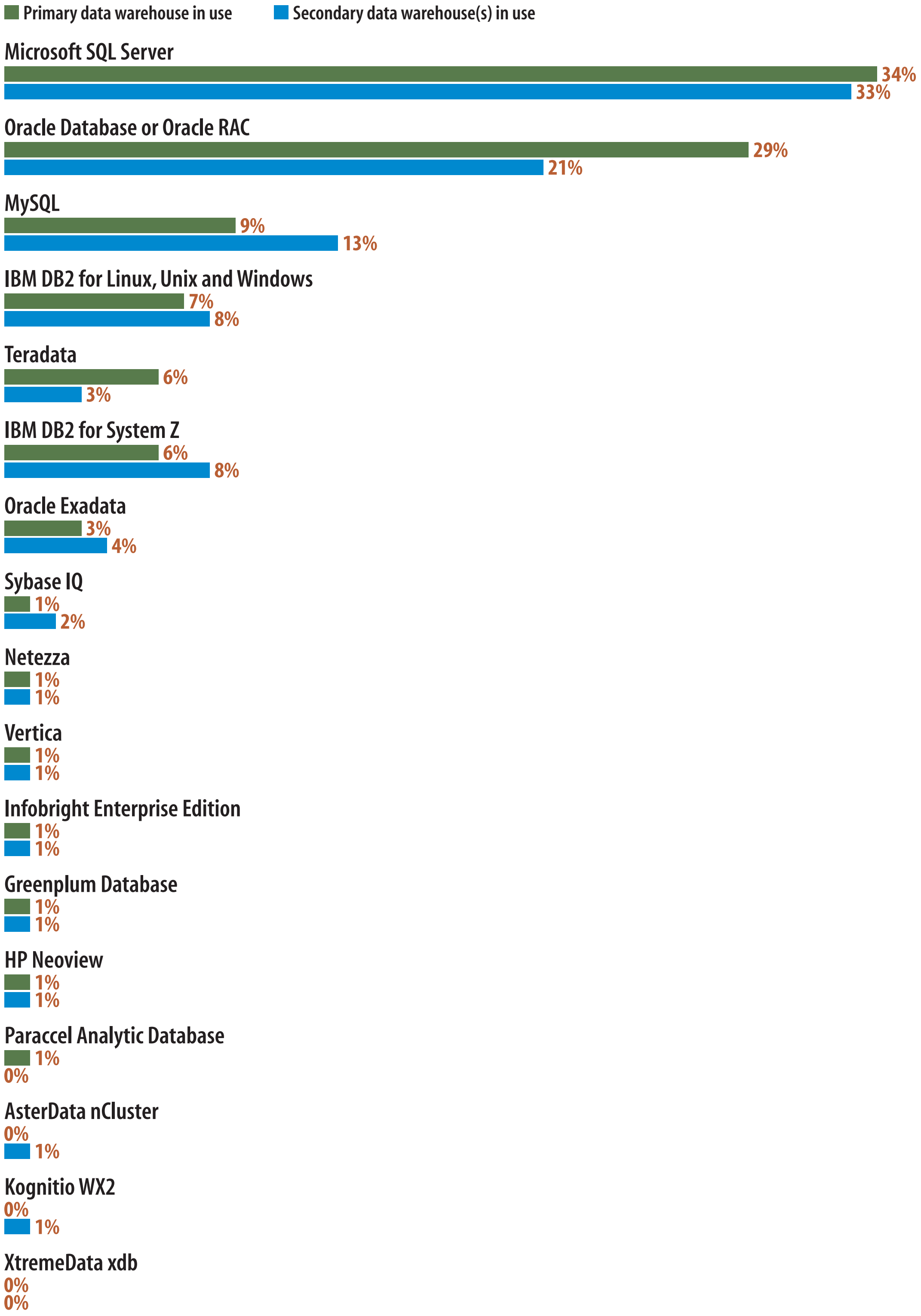
The advent of appliances is one of the most significant recent developments in the data warehouse field. As the term is used in connection with data warehousing, an appliance is a purpose-built, integrated hardware/software system that is engineered, configured, tested, delivered, priced and serviced as a unit, in marked contrast to the more widely used open systems model in which components—servers, storage, networking, operating system software and database software—are acquired independently and often integrated by the customer. The open systems approach has been favored by IT teams that want the freedom to independently select the components they prefer.

Many would say that Teradata's was the original appliance for data warehousing and was introduced in that form (though not by that name) decades ago. However, it was Netezza that in 2003 began to vigorously promote the term "data warehouse appliance." The company focused on rapid implementation, ease of administration, high performance via an MPP architecture

Figure 13

Future Use of Enterprise Data Warehouses

Within the next 12 to 18 months, which of the following do you expect to be the primary enterprise data warehouse in use at your organization? Which do you predict will be secondary in use?



Base: 580 respondents at organizations with one or more enterprise data warehouses

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

and aggressive pricing and met with significant success. Today, there are a number of data warehouse appliances on the market:

HP offers Neoview, an integrated MPP shared-nothing data warehouse product running on Intel-based HP hardware. HP does not specifically position Neoview as an appliance, although the product features the hardware/software integration featured in data warehouse appliance offerings.

IBM offers its family of IBM Smart Analytic Systems, fully integrated appliances with standardized configurations and comprehensive data warehouse capabilities. IBM's ISAS family includes the 7600, which runs on Power7 processors under AIX; the 5600, which runs on Intel architecture processors under Linux and Windows; the 9600, which runs the IBM System z operating system; and various other models. All ISAS products run DB2 as the database engine, employing a shared-nothing, MPP architecture.

Oracle now offers Exadata, the Sun Oracle Database Machine, an appliance in use for data warehousing and also for transaction processing. Inside an Exadata cabinet are eight servers running Oracle RAC and 14 servers running intelligent storage software developed specifically for Exadata. Oracle RAC runs the same way it does in other Oracle configurations, employing the shared database architecture that Oracle has been using for years, although Exadata V2 ships with Oracle 11gR2, a database that has been enhanced significantly over prior versions. There are eight cores in each server, so the database tier within the RAC cluster has 64 cores that can be employed to do most of the processing of database queries, including more complex joins and many other operations. However, an important part of query processing is offloaded to intelligent storage software running in parallel on the 14 servers. Here, scans can run in parallel, and because Exadata 2 has 11 disks per storage server in addition to solid state devices, the I/O bandwidth is considerable and the system delivers high performance on scan-intensive queries.

While it was introduced as a data warehouse platform, about 30% of Exadata systems shipped are in use either partly or entirely to support transaction processing. Exadata has also become popular as a consolidation platform running Oracle systems that were previously operating on many separate servers. Such consolidation can simplify administration and maintenance, lower cost and improve performance.

Teradata includes a range of appliances alongside its enterprise data warehouse platform, the Teradata 5600, including the Teradata 2580 and an extreme performance appliance, the Teradata 4600. Several startups, including Aster Data, Greenplum, nCluster, ParAccel Analytic Database, Vertica and XtremeData, have also entered the data warehouse market, employing highly parallel cluster architectures that are modular in the sense that capacity is readily expandable in units consisting of balanced configurations of servers and storage. Some of these companies provide an actual appliance (XtremeData does, and it is expected that Greenplum will, following its acquisition by EMC). Others deliver software because they want to appeal to customers who have their own preferred commodity hardware suppliers. All have been influenced by the appliance model and strive to simplify deployment.

Now, there is a difference between a true appliance and a system positioned as an appliance but actually supplied by multiple parties. In a true appliance, one party supplies a set of standard total system configurations that are tested, priced, sold, deployed and supported as integrated units. You don't deal with one party for the database software and another for the server—you deal with one party, period. Companies that want the benefits of an appliance need to inquire about the specifics of what is being provided by each prospective vendor and check into the experience of other customers.

Fall Into a Column

One much-promoted technique that has had some impact on data warehousing in the last few years is column storage. In column storage, the data in each column of a relational table is stored together. This is in contrast to the standard approach, in which the data in each row of a table is stored together.

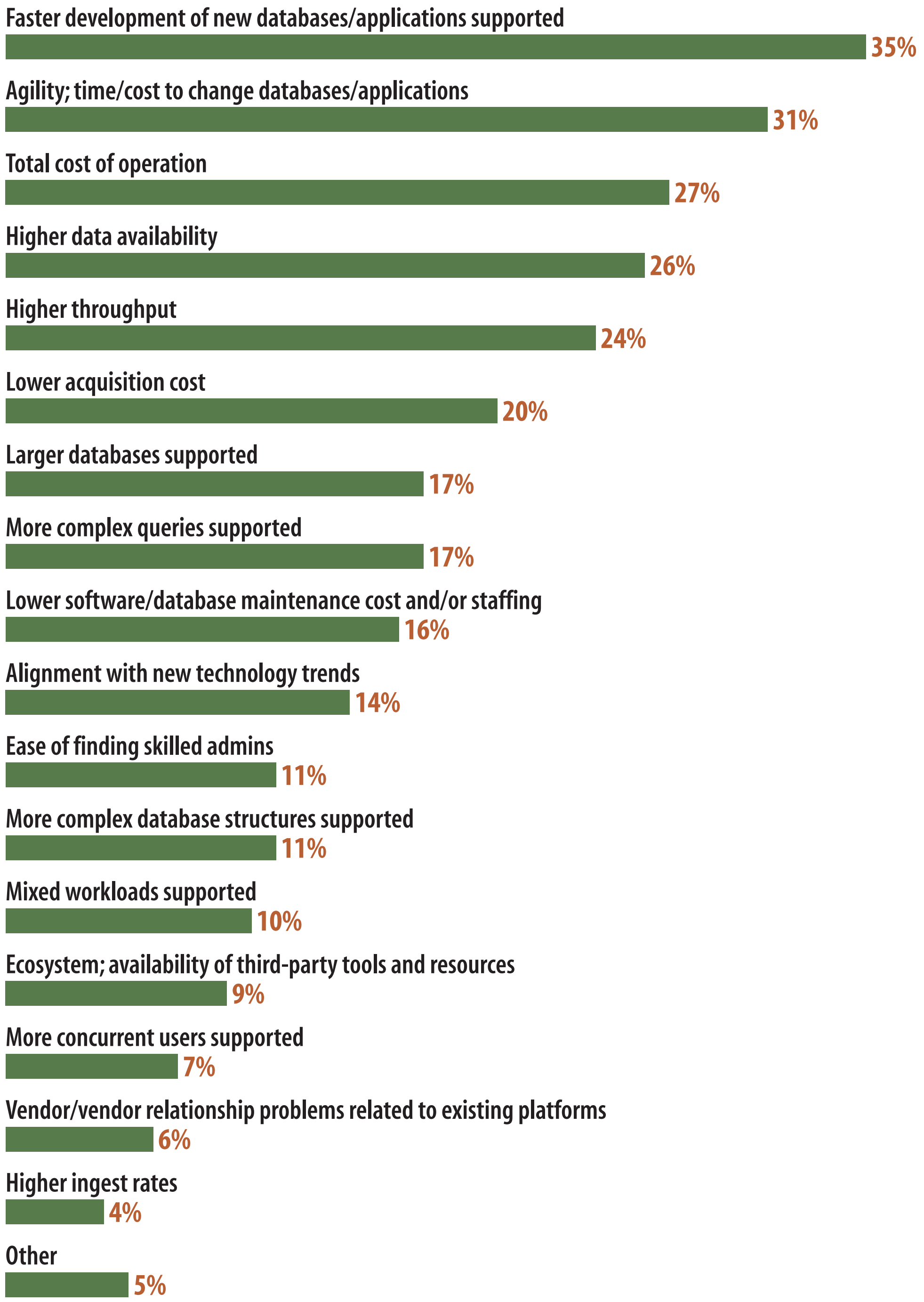
Proponents of column storage argue that the large tables in most data warehouses have many columns, typically hundreds, while only a few are referenced in each query. If you picture scanning a 100-column table to process a query that uses only 10 columns, you can see the point of the technique. In this case, you can do 10 times less work by storing the columns separately and reading only the ones you need. In addition, compressing data down to columns is more efficient than compressing across rows. Column storage enables column-wise data compression.

Sybase introduced the concept of column storage in the 1990s via its Sybase IQ, which exploits the technique for query processing, data compression and indexing. That is, the database

Figure 14

Factors Influencing Choice of Enterprise Data Mart

What are the top factors that influence your choice of enterprise data mart?



Note: Three responses allowed

Base: 422 respondents at organizations with one or more enterprise data marts

Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

engine can treat columns as indexes to rows. The Vertica Analytic Database, introduced in 2007, and the ParAccel Analytic Database, introduced in 2008, also employed column storage in combination with other techniques to provide data warehousing.

More recently, column storage concepts have been applied in database engines originally designed for row storage. Oracle Exadata employs hybrid columnar compression to automatically apply row storage to some tables; the Exadata software automatically determines when such a technique will pay off. Greenplum offers column storage as one option under its storage scheme.

Get Small

Data compression, a technique that's been around since the earliest days of computing, has taken on outsized importance in data warehousing over the last two to three years. In data compression, some unit of data is fed to a program that creates a smaller representation of that data, while preserving all the information contained. When the data is read back, the opposite transformation is applied. The compression/decompression process is invisible to the user, but system efficiencies result during the time the data is stored.

As a simple example of data compression, consider a database that is storing the names of the U.S. states in which customers live. A natural way to represent such data would be to spell out the state name in plain English, so customers who live in Boston would have the value "Massachusetts" stored in the state column. As there are 50 states, a simple compression scheme would assign a number between 1 and 50 to each state; store the number in place of the state name; and each time the state value was retrieved, substitute the actual name for the number.

Database products that support data compression do all this automatically and apply a variety of transformations depending on the type and frequency of data values. The process can become quite complex and works better on some types of data than others. In some systems, data compression is an option under the control of the database administrator; in others, the system automatically decides when it will pay off.

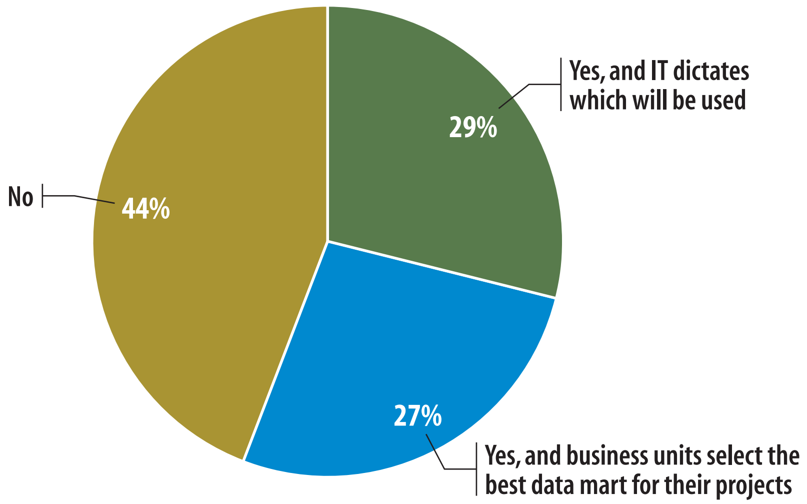
While data compression was primarily conceived to save on storage space, the larger present day motivation is to increase system performance and reduce overall system capacity requirements. Large compression ratios are achievable on some types of data with some products—we've seen vendors claim ratios as high as 10 times on active data and 50 times on archival data. In reality, most users experience somewhere between 2 times and 5 times, which is nothing to sneeze at. Scanning data from disk storage, a typical database process will realize somewhere around 100 Mbps, so reading a TB of data will require about 10,000 disk seconds. Reading in parallel from 100 disks with perfect scalability will reduce the time to 100 seconds. In many settings, 100 seconds is too long to take to answer a question. But, if the data is compressed by a factor of two, and assuming decompression adds no elapsed time, then the same data will read in 50 seconds, reducing the response time by half.

Because processor capacity has been rising much faster than disk I/O capacity, during the last few years database vendors have built increasingly effective compression into their products. Netezza employs a field-programmable gate array—a relatively inexpensive, specialized processor—to do decompression at a rate that keeps up with disk I/O. The result of this approach is that the main processor never even sees the compressed data.

Figure 15

State of Enterprise Data Mart

Do you have one or more enterprise data marts?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Oracle has built hybrid columnar compression into its intelligent Exadata storage cells, with a similar effect; the processors in the Oracle database tier see only the uncompressed data. And IBM, after providing data compression in DB2 some time ago, has recently released an optional feature informally known as "deep compression," said to result in large compression ratios.

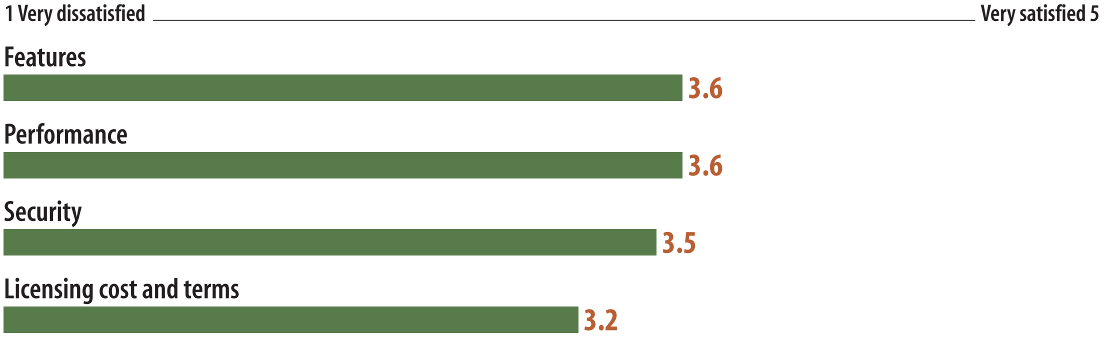
Building data compression into the data warehouse can often result in large savings in total system cost. In an I/O-bound, query-intensive system, a 2 times data compression ratio can mean a 50% reduction in the total cost of the system. But there is a caveat: Compression ratios do vary with the data used. Since many vendors now quote prices based on certain assumed compression ratios, CIOs need to be very careful when comparing. We recommend using your own data to test the compression level actually achievable in practice and factoring this into your interpretation of each price quote. Don't assume the ratio will be the same on different products—you need to test with a large sample of real data.

Also keep in mind that some compression features are designed for archival data only. Since archival data is rarely read, decompression does not have to be as efficient as it does for frequently accessed data. Higher compression ratios can be achieved for archival data, but these

Figure 16

Satisfaction With Enterprise Data Mart

Overall, what is your level of satisfaction in the following areas as it applies to your current enterprise data mart? Please use a scale of 1 to 5, where 1 is "very dissatisfied" and 5 is "very satisfied."



Note: Mean average ratings
Base: 422 respondents at organizations with one or more enterprise data marts
Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

should be taken into account only when configuring archival systems. Compression is only one aspect of system capacity calculations, but correct interpretation of data compression rates is particularly important because mistakes can lead to gross underestimates of the real cost of a system.

Extreme Analytics

Over the last few years, we've come to view a certain set of analytic problems as falling into a separate category, dubbed "extreme analytics." Usually included in this bucket are situations where data volumes are very large—hundreds of terabytes to petabytes—and analysis requirements are intensive. Often, the analysis is clumsy or impractical to perform entirely in SQL, a nonprocedural language, and is attacked primarily with routines or functions written in a procedural language such as Java.

In our survey, 48% of respondents say they view analytic databases as a separate category from data warehouses or data marts. Further, 67% of these respondents say that they have analytic databases and applications that are independent of their data warehouse/data mart environments. Respondents who express interest in analytic databases cite a somewhat different set of top factors in platform selection: faster development, higher throughput and alignment with technology trends. This in contrast with the top two EDW concerns, TCO and data availability.

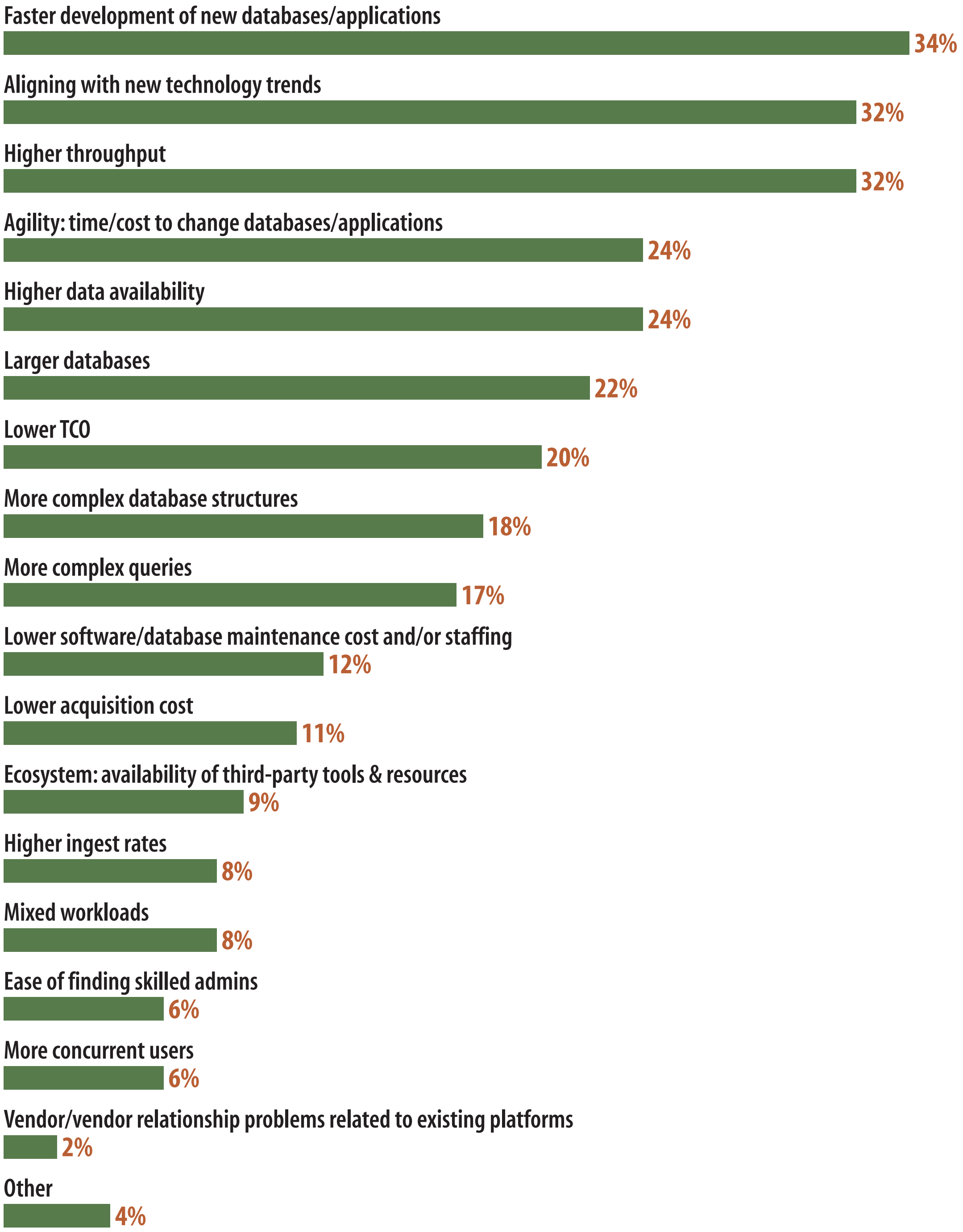
Some usage scenarios in analytics are indeed quite different from the focus in data warehousing. Data warehousing, at its core, is about leveraging data over multiple uses and, often, over a long period of time. It features making an upfront investment in carefully defining, modeling, cleansing and integrating data so that it can be applied to a variety of different purposes, typically over some years.

While analytics features a range of scenarios as well, some of which mesh with a data warehousing approach, analytics also includes applications where the data is used by a small number of people—sometimes, one person—for a short period of time. It's often used in scientific laboratories, engineering applications, some areas of financial analysis and other settings where it's common to suddenly receive huge volumes of data, often petabytes, that must be analyzed quickly but not retained. Or, it may be important to analyze a petabyte of new data every day to select a terabyte subset (1/1,000th) to retain for longer-term use.

Figure 17

Factors Driving Interest in Integrated Analytical Database Platform

What are the top factors driving your interest in an approach that integrates your analytical database platform with MapReduce, Bigtable and/or Hadoop?



Note: Three responses allowed
Base: 102 respondents interested in an integrated analytical database platform
Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Some commercial products are aimed at analytics. XtremeData, for example, offers an analytic data appliance designed specifically for this situation. Teradata also has an Extreme Data Appliance designed to economically handle very large data volumes (currently up to 50 PB) and an Extreme Performance Appliance aimed in part at intensive data analysis requirements.

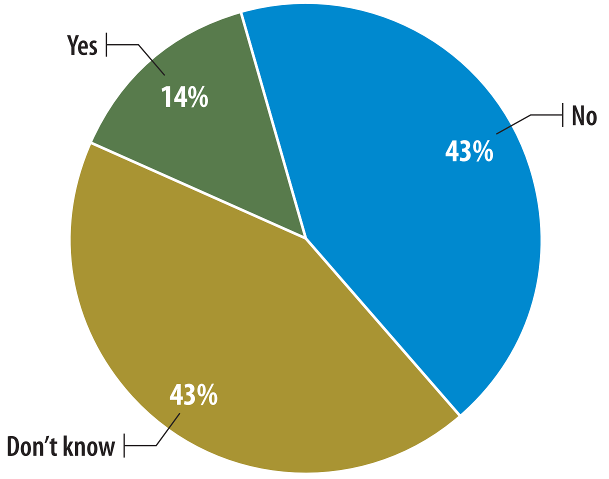
In addition, some data warehouse vendors have integrated technologies into their data warehouse engines to respond to analytic requirements for data that has been, or will be, incorporated into the data warehouse environment. It has long been the practice in many enterprises to extract data from the data warehouse environment, transport it to a separate server, and there apply analytical tools for some type of analysis that was not readily accomplished in SQL. One of the most commonly used tools in this case is SAS and, indeed, many companies have a separate SAS server or infrastructure maintained exactly for this purpose.

A key problem with that scenario is this: If the data already resides in the data warehouse, and

Figure 18

Interest in Integrated Analytical Database Platform

Are you interested in an approach that integrates your analytical database platform with MapReduce, Bigtable and/or Hadoop?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

the volume of data to be analyzed is considerable, then it's problematic just to move the data from the warehouse to the analytical server. In addition, if the analytical tool is not capable of high-performance processing, as in a highly parallel architecture, then it can take a long time to do the analysis. One solution is to perform the analysis in place on the data warehouse, exploiting its highly parallel architecture. Teradata has delivered capabilities for performing SAS routines in place, and Netezza has announced such capabilities as part of Release 6 of its software. If you expect to perform a significant number of analytics operations, ask current and prospective vendors about their plans in this area.

Big, Fast and Open

A key requirement in extreme analytics is to apply enormous amounts of computer power to analyze huge volumes of data, economically. Partly in response to these requirements, new technologies have emerged from Google, Yahoo and other companies involved in very-large-

The Time and the Place

"When" and "Where" have been fundamentals in news reporting since before we had computers, and they have been fundamentally important aspects of data analysis as long as we have had data. Remarkably, however, database products have not helped users deal with time and location data as much as they could, even though many database products have long featured data types for time, date and location.

That needs to change, however, because the last few years have seen an explosion in the use of mobile devices, including those for GPS; widespread use of inexpensive sensors and cameras; and other developments that have flooded computer systems with time and location data. This has prompted interest in handling large volumes of such data efficiently and enhancing database capabilities for defining and enforcing time and place semantics. Oracle, DB2 and other database products now feature support for time and location data. Teradata says it will provide new capabilities for temporal and location data in its new release, due this fall, including enhancements in the performance of time- and location-based queries and built-in support for transaction time (the time when a fact is stored in the database) and valid time (the time when an event occurs in reality, or the "reality" modeled in the database).

If this type of data is important to your business, ensure you ask your database vendors how they handle it.

scale Internet businesses. These have been embodied in Hadoop (hadoop.apache.org), an Apache Foundation open source project. Key elements of Hadoop are:

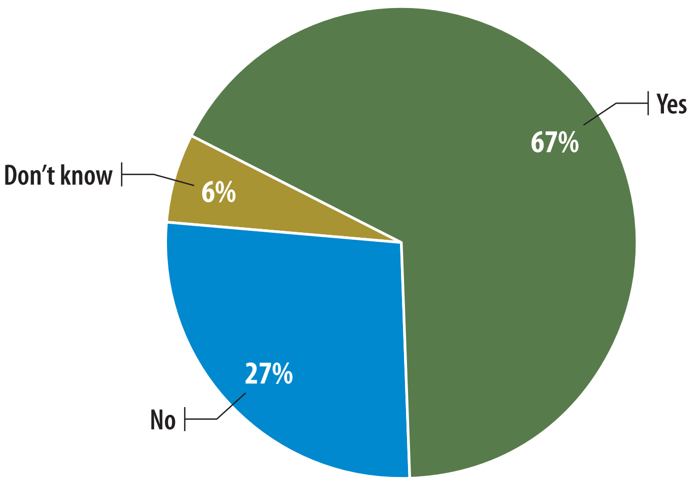
- **HBase:** A scalable, distributed database that supports structured data storage for large tables;
- **HDFS:** A distributed file system that provides high-throughput access to application data;
- **Hive:** A data warehouse infrastructure that provides data summarization and ad hoc querying; and
- **MapReduce:** A software framework for distributed processing of large data sets on compute clusters.

Approximately 80 projects using Hadoop are listed at the "Powered By" section of the Hadoop Wiki (http://wiki.apache.org/hadoop/PoweredBy). As an example of a very large project, Yahoo runs Hadoop on more than 36,000 computers. The largest Yahoo cluster described has 4,000 nodes and is used for research on advertising systems and Web search.

Figure 19

Use of Independent Analytic Databases and Applications

Do you have analytic databases/applications that are independent of your data warehouse/mart environment(s)?



Base: 360 respondents who consider analytic databases and applications separate from data warehouses and data marts
Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

There is also commercial activity coalescing around this open source software. Cloudera offers an enterprise distribution of Hadoop that includes enhancements for increased stability and management tools, and IBM began a research initiative in this area about four years ago.

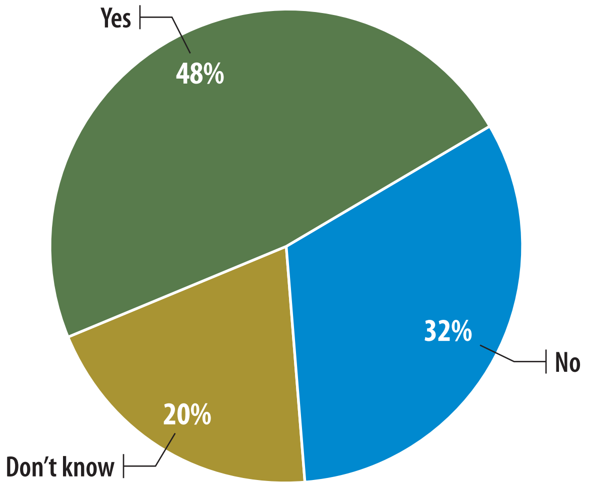
It's now offering its own enhanced distribution of Hadoop via a services initiative named BigInsights (www-01.ibm.com/software/data/infosphere/hadoop/) that's aimed at helping customers who want to use Hadoop-related capabilities for large-scale analysis.

Hamid Pirahesh, an IBM fellow and the leader of IBM's research program in extreme analytics, says companies across a range of businesses are finding that they must store and analyze enormous volumes of data, often not in structured, tabular form, from a large and rapidly expanding array of sources. Common examples include records structured as key-value pairs, documents such as blogs and e-mail messages, and data that may have a graphical structure.

Figure 20

Analytic Databases Separate Category From Data Warehouses and Data Marts?

Do you consider analytic databases and applications to be a legitimate separate category from data warehouses and data marts?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Increasingly, only a portion of this data ends up in data warehouses. Much of it could go into a Hadoop repository and be analyzed there using MapReduce.

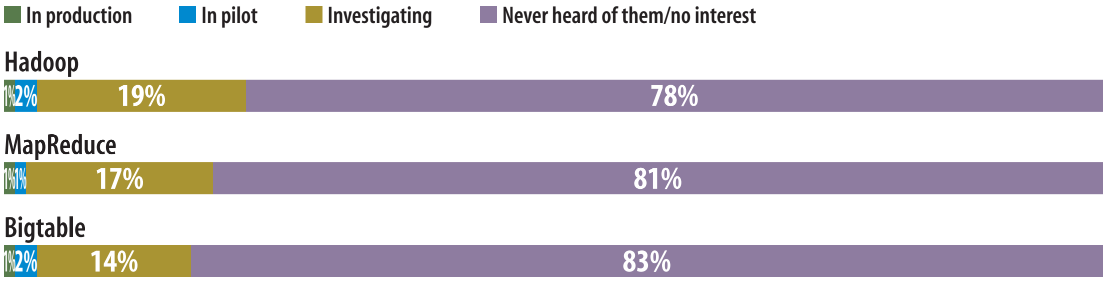
However, many organizations do not want to write programs that make MapReduce calls to analyze this data, according to Pirahesh. As a result, IBM is developing higher-level languages that can be used by analysts to access and analyze data in Hadoop. Pirahesh says that many IBM customers are pursuing the use of large-scale Hadoop environments for analytics but place a high priority on making such data and capabilities available to existing users of business intelligence, the data warehouse and commercial analytic environments. Many of these users cannot program in Java or else are accustomed to working in other environments, such SAS or SPSS. Also, the data in Hadoop cannot be used in a vacuum—there must be a capability to integrate it with high-quality data in the data warehouse and with master data.

Teradata has announced an interface using Cloudera to facilitate moving data back and forth between the Teradata data warehouse and a Hadoop environment. Earlier, Greenplum and Vertica enhanced their data warehouse products with connections to programs running in Hadoop environments. Greenplum describes a scheme in which data-flow procedures defined to Greenplum can perform processes, which perform some steps in the Greenplum database and others in the Hadoop environment.

Figure 21

Use of Technologies for Analytics Applications and Databases

What is your degree of use or interest in the following technologies for analytic applications and databases?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Aster Data has taken a different approach, in which MapReduce calls can be embedded in SQL queries and performed in place on the database in Aster Data nCluster. The Aster Data MapReduce implementation is in use at MySpace and other major Web-based businesses.

Keep It Safe

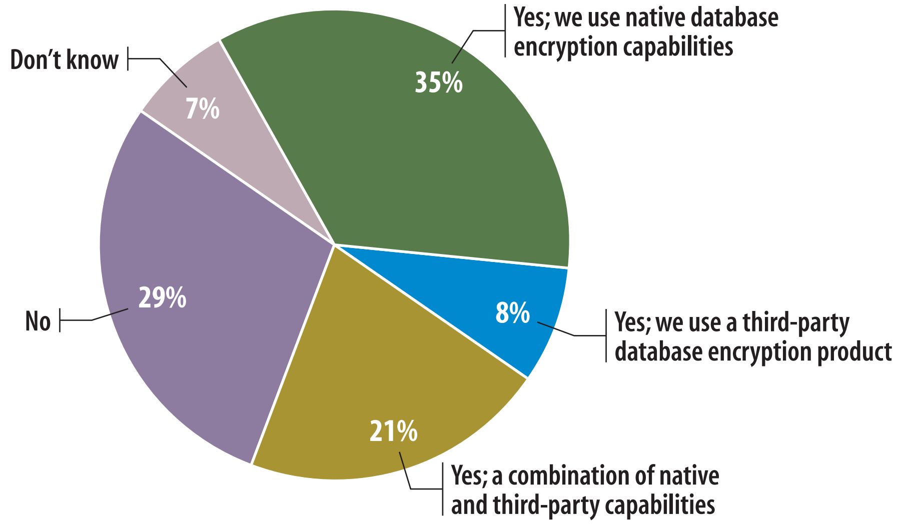
All the analytics power in the world won't help you if you ignore database security or focus only on external boundaries. Of our poll respondents, 64% are using encryption on their databases. If the encryption is applied outside of the database—say, at the disk drive level—this is primarily a precaution against theft of the media or the data when it's outside of the database system.

For example, if someone copies the database files using an operating system utility and then accesses them with tools other than the database itself, this type of encryption should defeat efforts to use the data. However, when data is accessed in place through the database system,

Figure 22

Use of Encryption on Databases With Sensitive Information

Does your organization use database encryption on databases that contain sensitive information?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

perhaps with a stolen user account and password, the encryption may not help, since the data is decrypted for such authorized use.

Nonetheless, database encryption is a fundamental precaution for data protection. There are many environments where it would be easier to surreptitiously copy a file than it would be to steal a database account and password. Some users hesitate to use encryption for fear it will complicate recovery from system failures and/or disasters—they worry that the means to decrypt their data will somehow be lost, compromised or unavailable at a crucial moment. This risk must be weighed against the possibility of a large-scale loss of sensitive data.

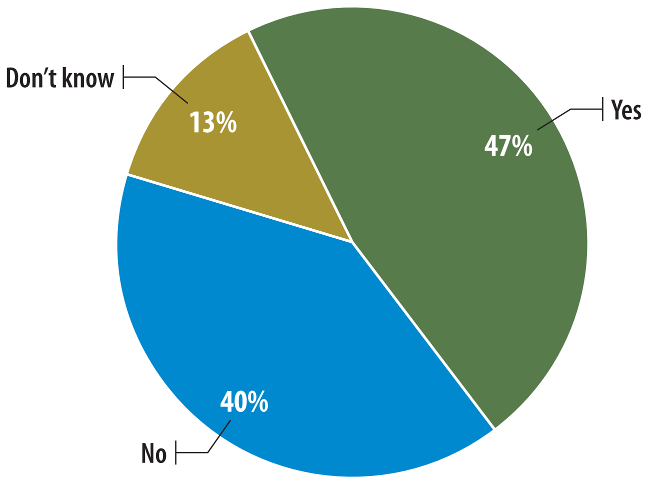
In general, we would like to see database encryption more universally applied; it is one practical tool that will increase the cost and difficulty of data theft.

Our respondents seem to do better at some of the procedural elements of data security. For

Figure 23

Use of a Database Firewall

Does your organization use a database firewall?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

example, 74% have transaction logging enabled for all databases containing sensitive information, which does provide a place to start when investigating breaches. And 70% say their organizations perform database security assessments, which can identify weak areas before problems occur and assist in directing resources where needed—oddly, however, just 36% of those performing these assessments were able to name the security assessment products in use.

As we discuss in our Dark Reading Tech Center, at www.darkreading.com/database_security/index.jhtml, database security needs to be a priority. If you're not among the 37% with a defined procedure for conducting forensic investigation after a database compromise, that's a fine place to begin.

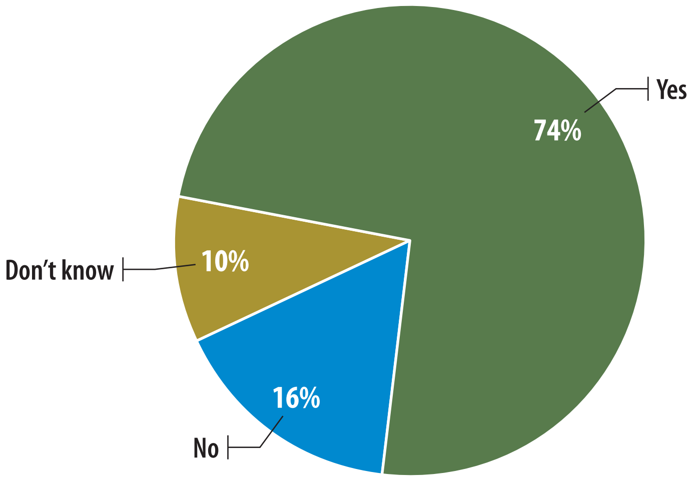
Putting It All Together

Actually realizing the database management benefits we've discussed—even a significant subset

Figure 24

Transaction Logging Enabled on Databases With Sensitive Information

Does your organization have transaction logging enabled for all databases containing sensitive information?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

of them—will yield business value well above the investment in database products, staff, security and data management programs.

Nevertheless, it is often difficult to capture or measure these elements of value. When it comes to establishing budgets, we tend to focus on either acquisition cost or total operating cost for the database management platform.

Total operating cost: A typical breakdown of total operating cost for a database platform includes:

1. Platform acquisition cost (may be bundled into a single item)
    - Database software, including utilities
    - System software
    - Servers
    - Storage
    - Network
2. Platform support fees (same elements as #1)
3. Platform upgrades (same elements as #1)
4. Support staff
    - Database administration
    - System administration
    - Storage administration
    - System operations
    - Security
5. Environmental Factors
    - Power
    - Space
    - Cooling

Comparing these costs to the areas of value we've discussed is an interesting exercise. The efficiency with which a platform performs or the economy with which it scales to handle more data directly affect the overall cost. The same can be said of platforms that require significantly less support staff or space and power.

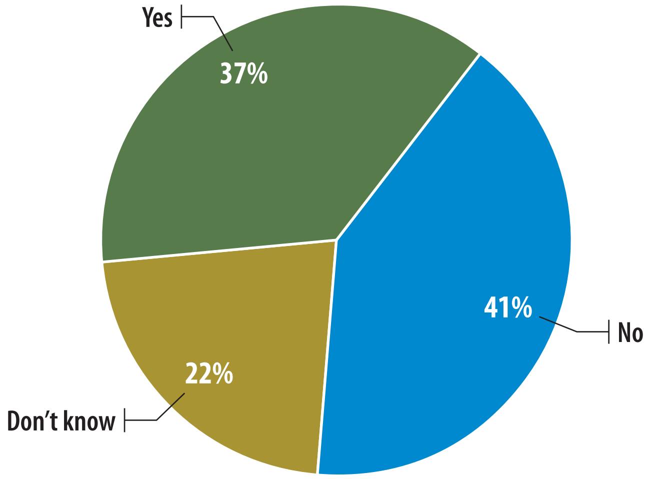
Similarly, the manageability of a platform can drastically affect the support staff requirements, which are often the largest cost factor and the biggest management challenge in the long run.

In fact, many of the elements of value from a successful database management program apply to other areas of the business or IT operation and are much larger in their impact than the database cost. In our experience, most businesses spend at least 10 times as much money—and employ 10 times the staff—in application development and maintenance as on database management. We've seen many cases where as much as 40% of application development costs are devoted to sourcing, gaining access to and using data. Thus, a database management program that eases and standardizes application access to data could cus this cost in half, and most like-

Figure 25

Defined Procedures for Conducting Forensic Investigation

Does your organization have a defined procedure for conducting forensic investigation after a database compromise?



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

ly pay for itself, even before taking into account the value of more agility, better information and/or better decision-making in the business.

Unfortunately, few enterprises measure cost/benefit in this way. It is therefore an ongoing challenge to effectively link business value to better database investment decisions. Our advice here:

- Get the business objectives right, and then map out the data management elements needed to support them. Misunderstanding the problem to be solved—and consequently choosing the wrong platform—is a common cause of project failure.

- Think through the associated technical database requirements. Focus on managing data well by breaking down silos, integrating data and leveraging it across the enterprise and over time. Though databases may behave like commodity products in many projects with routine requirements, they differ drastically in their performance, scalability, manageability and TCO when requirements are more demanding. End up on the wrong platform, and you won't have the scalability or performance chops to deal with database and workload growth. Fail to contain complexity or foresee rapid change in your data and applications and you'll end up spending a fortune on staff and consultants because it has become too difficult to keep the database intact and operating correctly.

- When doing budget analysis, focus on the total cost of operations. License fees are significant, but by no means are they the biggest cost or the biggest lever with respect to business value in most situations. With strategic databases investments, weigh business value and business risk. Incorporate new, low-cost and open source technologies judiciously. And keep in mind that you can save much of what you spend on application development with better data management.

- Plan a database architecture that can handle rapidly escalating scale and complexity. Base your platform decisions on your business objectives and their technical implications. Take special care to capture requirements for scale, schema complexity and query complexity likely to develop over the next two to three years.

- Choose database technologies that facilitate rapid application development and ease of change while minimizing the cost and effort of system administration. Database platform decisions are still fraught with risk, especially in data warehousing and analytics, but enterprises need, or soon will need, a cost-effective way of dealing with extraordinary volumes of data and intensive analysis. Strive to integrate your extreme analytics system with the rest of your IT environment, especially the business intelligence/data warehouse portion. Extreme analytics systems are not going to produce maximum value if they are implemented as one more island of information.

- Put processes in place that will enable you to keep architectural complexity and data latency in check as business needs evolve and data volumes grow.

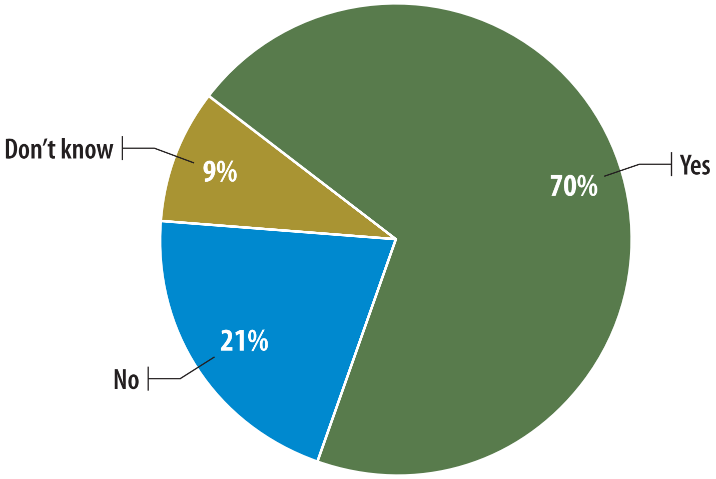
Parting Wisdom

IT leaders are under continuing pressure to control costs, even as data and workload volumes skyrocket and user expectations become less forgiving. Application performance and data availability have to be good—often all the time—and able to scale up as the business grows. All these pressures converge at the database and create a climate where neither high cost nor sys-

Figure 26

Perform Database Security Assessments

Does your organization perform database security assessments?

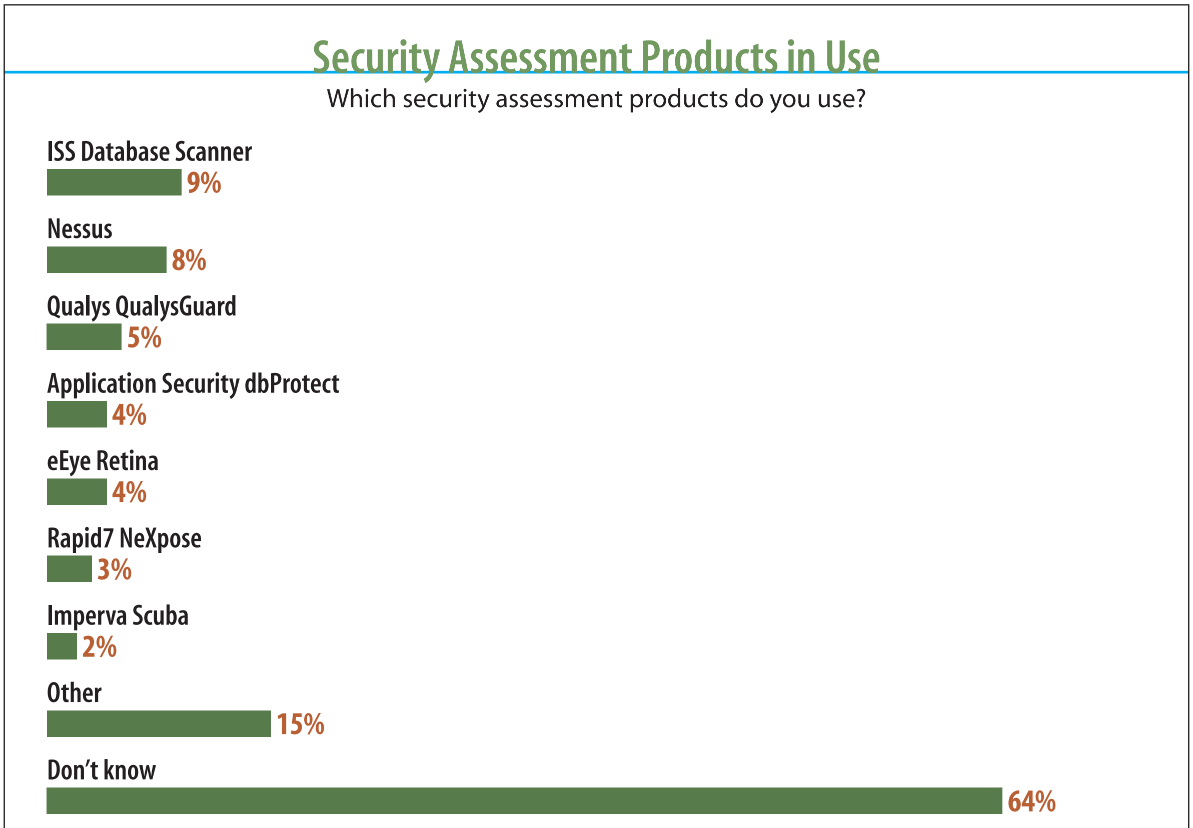


Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

tem failure are allowed. This combination of forces is causing users to cast around for new approaches to growing database requirements, and this is evident in our data.

The database systems we have known for a long time, from Oracle, Microsoft and IBM, are the most widely used and continue to improve rapidly along multiple fronts. Teradata is in a similar position in data warehousing. But CIOs are demanding more: Systems that are easier to implement, easier to manage, better performing and less costly. Established vendors have reacted with a range of innovation: IBM has its Smart Analytic Systems. Oracle has expanded its architecture with Exadata, increasing I/O parallelism and building intelligence into the storage

Figure 27



Note: Multiple responses allowed
Base: 528 respondents at organizations performing database security assessments
Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

layer. Microsoft has developed Microsoft SQL Server 2008 R2 Parallel Data Warehouse, which is set to be released by the end of this year, with similar objectives. HP is in the game with NonStop and Neoview. Teradata's appliance family is worth watching. Meanwhile, there have never been more startups to choose from. Netezza has shown it is possible to disrupt this mature market, make money, go public and continue to grow. A half dozen or more upstarts are going for a piece of the data warehouse and/or analytics market. Meanwhile, an entirely new segment is emerging to meet extreme analytics requirements. Massive systems are in operation at Google, Yahoo and many other companies. IT executives in hundreds of other businesses are wrestling with decisions on where these systems fit in.

With such a dynamic picture, decision-makers have to play it smart. Though some database startups succeed, many others never really take off. It pays to continue to use proven (albeit with high license fees) systems while experimenting with and evaluating new technologies and approaches. And, don't underestimate the advances being made by established vendors.

In other words, don't chase the cool factor—you have to be guided by your own vision, requirements and constraints. And you have to devise meaningful ways to test new technologies and products before relying too much on them. With databases, validation is always tricky; because problems tend to surface later, quick little benchmarks don't tell you enough. Above all, avoid the trap of the superficial test. Don't let the vendor design a short demo to prove your requirements will be met. Test at realistic levels of complexity and scale before making decisions. When you check out reference sites, make sure you find some that are solving the problems you will face in the next few years.

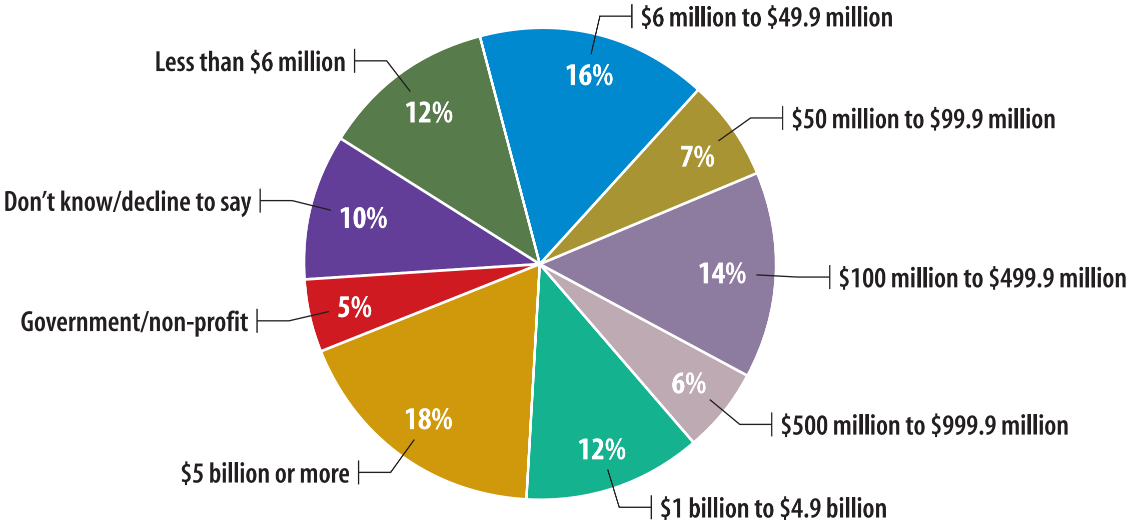
In general, think through your database requirements, test for real capabilities to meet them, then keep your eye on the risks as you work through pilot programs and early applications. Manage carefully as the scale grows large. Keep these basic principles in mind, and you'll find you can benefit from the both the established products and the rapid changes in the database field.

Appendix

Figure 28

Company Revenue

Which of the following dollar ranges includes the annual revenue of your entire organization?

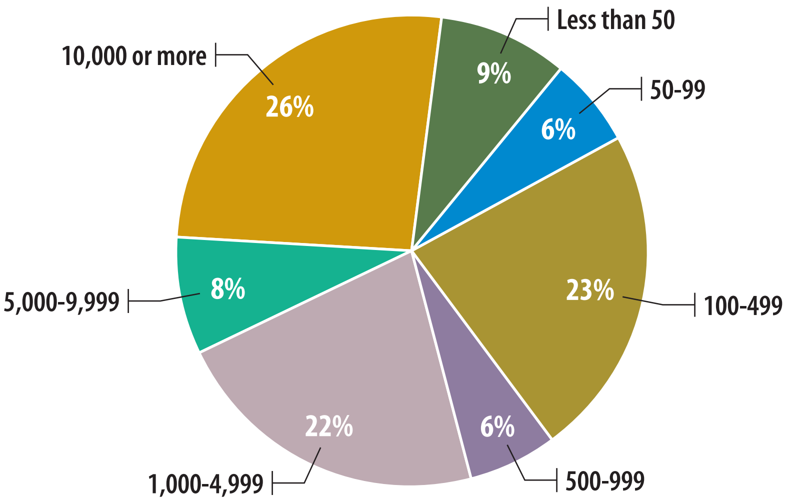


Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Figure 29

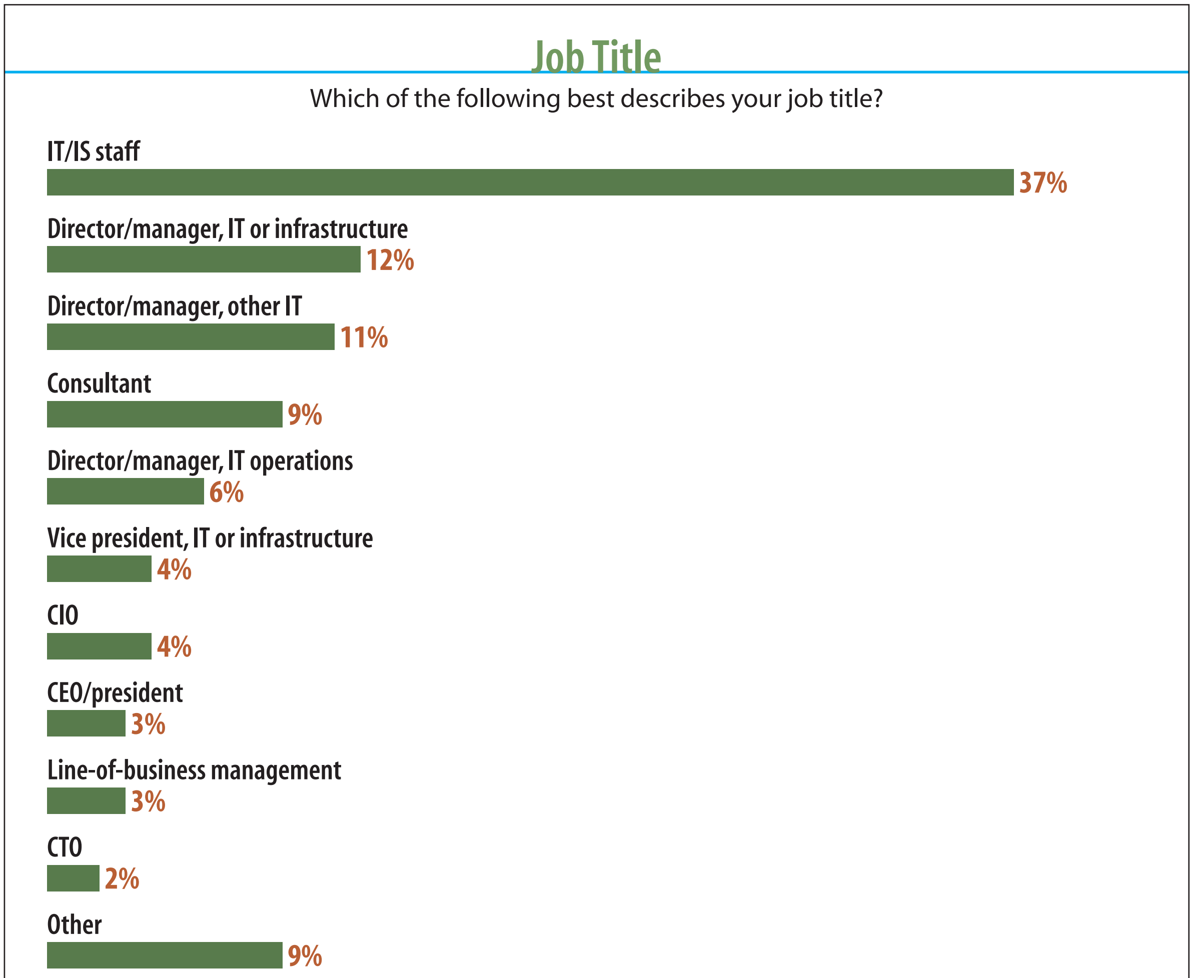
Company Size

Approximately how many employees are in your organization?



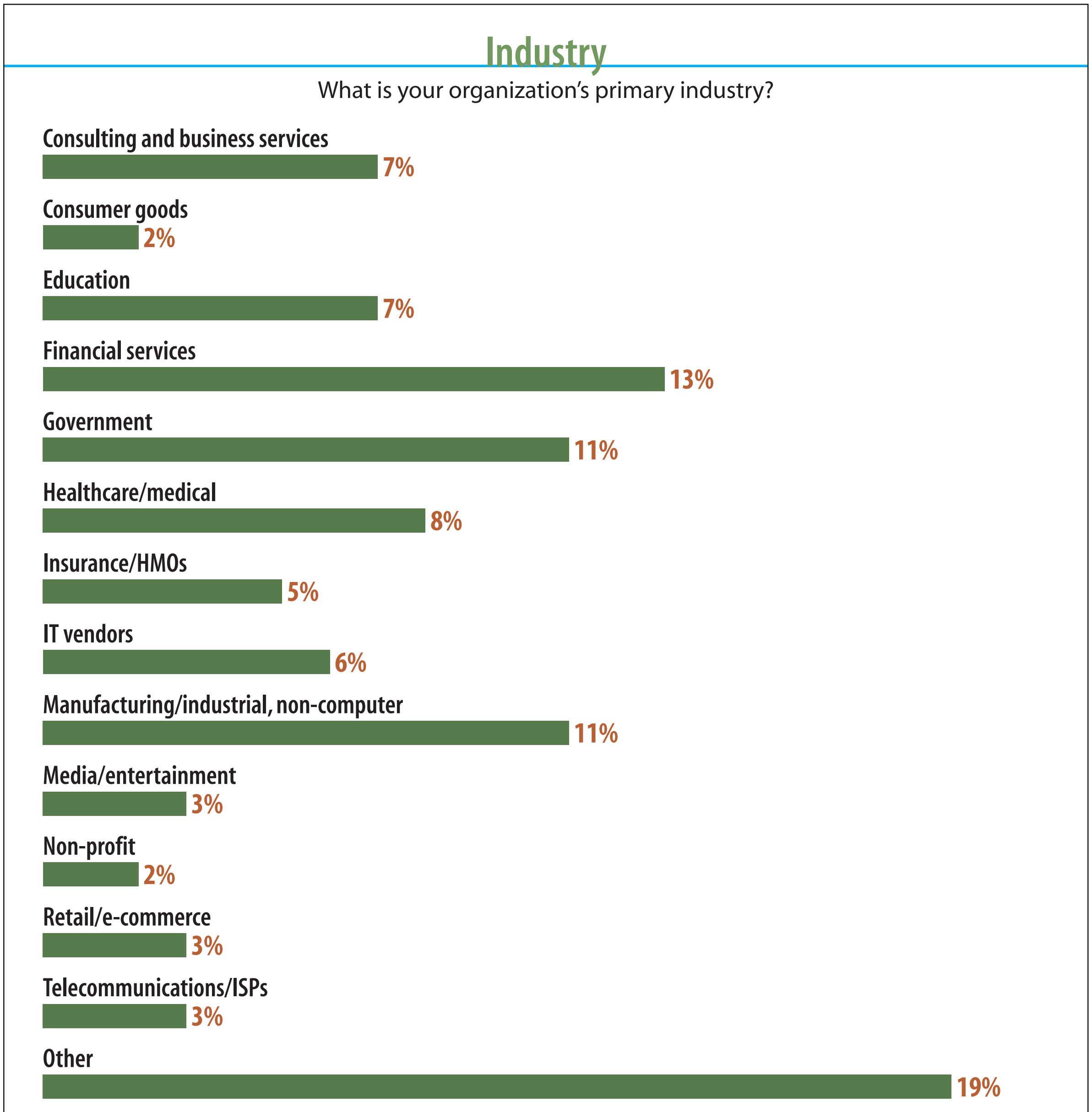
Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Figure 30



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010

Figure 31



Data: *InformationWeek Analytics* 2010 State of Database Technology Survey of 755 business technology professionals, August 2010