Customer Experiences Migrating Oracle BI Workloads to a Teradata Solution

2011 Technology White Paper

Third Nature

TABLE OF CONTENTS

Introduction

Why are so many people going through the pain of migrating to a new analytic database? Moving a data warehouse from traditional general-purpose databases like Oracle to an analytic database is not an easy exercise. The database is at the heart of the environment, making it the most difficult piece of the system to change. It has links to everything else: ETL, business intelligence (BI) tools, data mining tools, data transfer to other applications, monitoring, backups and staff trained to manage it.

Database sizes increased significantly across all industries during the past few years, but performance at the sub-terabyte scale is still a challenge for most organizations. A successful migration project requires understanding the true nature of the problems being faced.

This paper is the result of a series of IT interviews exploring what drove organizations to move their BI environments away from traditional databases like Oracle. While performance and scalability are at the root of most problems, there are four basic scenarios that drive the move to an analytic database.

People choose different approaches to migrate to a new environment based on these four scenarios. These approaches deliver different architectures and have different project plans and assumptions associated with them. This paper identifies the common scenarios, the approaches taken to successfully migrate to a new platform, and the lessons learned during the projects.

Why are People Using Analytic Databases?

People describe many reasons for needing to move away from their database: growth in data, increased use, the need to provide a single consolidated set of information, and evolution of the BI workload to meet new needs are all cited. The core decision is always motivated by concerns about performance and scalability at a manageable cost.

Performance isn't strictly a 'big data' problem involving hundreds of terabytes of data. Most people interviewed were having problems with less than four terabytes, some at sub-terabyte scale. If the inability to get good performance is a problem over moderately large data sizes, why not buy more hardware to solve the problem? Is it necessary to change the database when hardware gets faster and cheaper every year?

The answer for many people is "yes" because they report diminishing returns as they tune the database and add more hardware. After several iterations, they approach a performance ceiling that is difficult to raise. It's at this point that hardware and people costs become a concern.

"We looked at what it would take to grow our Oracle environment. The hardware and license costs almost equaled the purchase price for Teradata, with no guarantee we wouldn't hit the same problem in a year," said one IT manager. "If you've tried all the usual options and are looking at a big investment in the Oracle environment, it's time to consider analytic databases designed specifically for the workload and compare the tradeoffs."

Oracle's Exadata was an appealing alternative to some customers because it would allow them to stay with a database they knew. The problem was that the Exadata change could be as disruptive as moving to a completely different platform. And it might not be any cheaper than trading up to an analytic database such as the Teradata® Database.

Cost is a big factor for many people, but it isn't the single most important one. The real factor is price-performance and the price attached to scaling up. When comparing traditional and analytic databases, the choice is often whether to pay to be at the ceiling of traditional database performance, or to be at the floor of analytic database performance.

One architect described trying many solutions over a two-year period to resolve performance. "We tried a lot of things. We brought in database performance consultants. We tuned. We bought new hardware. The biggest problem was the conflicting advice we would get from different consultants."

This is a common complaint. There are many Oracle generalists available. There aren't many who understand how to tune and manage specifically for a BI workload, particularly across different hardware environments. One of the advantages of analytic databases like Teradata is a focus on only BI workloads, simplifying the hardware environment and the task of performance tuning.

This architect's attempt to resolve problems is reflected in others' experiences as shown in Figure 1. The chart shows answers to a question about what people have done to address performance problems in their BI environment, irrespective of database type.
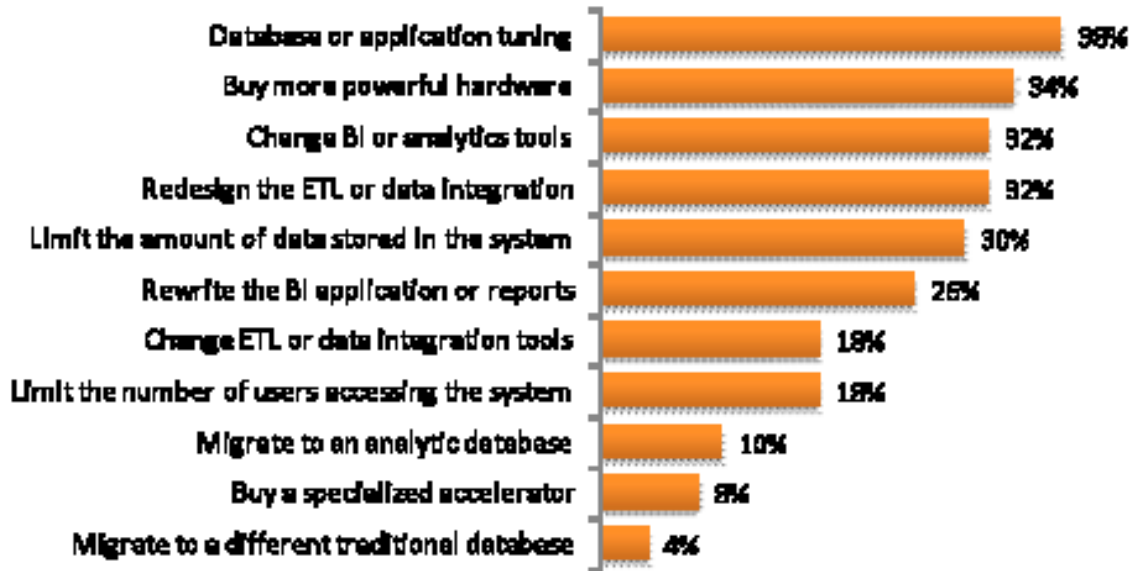
Figure 1: Answers to "What did you try in order to resolve performance problems?"

Figure 1 shows that tuning and buying more hardware are the most frequent practices. What is surprising is the willingness to change, redesign or replace every other tool in the BI stack before replacing the database. This includes practices like limiting user access or purging useful data to shrink the database, actions that can reduce the business value and usefulness of the warehouse.

When faced with making sacrifices like these, it's worth looking at the experiences of other organizations to see what patterns emerge. Understanding the common scenarios that drive people to migrate to an analytic database will help when deciding if this is an option for you.

A successful migration project depends on identifying the specifics of the situation and finding the approach that fits your needs best. There are multiple approaches, so it will be important to understand the tradeoffs between them. The rest of this paper describes common scenarios faced by organizations, the approaches they took to migrate their data warehouse, and the lessons they learned from the projects.

Common Scenarios Driving Data Warehouse Migrations

The reasons organizations give for moving to an analytic database fall into four easily recognizable categories outlined below.

Performance

The single most common complaint is trying to run the existing workload in a reasonable amount of time. The warehouse is either stable or growing at a predictable rate, yet performance is slow or is unpredictable.

When surveyed, BI managers reported the areas of poor performance shown in Figure 2. Query performance dominates complaints. This is not a surprise because query is the most visible element and affects the largest number of people. It's also the most difficult to diagnose because many technology and design factors can affect query response.

Figure 2: Performance complaints reported. **Note:** *Numbers add up to more than 100% because people reported more than one problem.*

One third of organizations have trouble loading their data on time, leading to complaints about data availability. Most people who had this complaint also had associated interactive response problems, indicating that their systems are running at full capacity. Others worked around data integration problems by redesigning ETL or running it more frequently to reduce the peak workload.

Query performance complaints are growing as more organizations roll out dashboards, scorecards and visualization tools that require interactive response speeds. Several people said that adding dashboards or scorecards caused more performance complaints because the additional workload slowed the system or because tuning to give interactive applications faster response lowered the priority of other BI work.

A related problem was the need to load data more frequently to support intra-day dashboard needs. This slows the database because both data loading and querying happen at the same time and can compete for resources.

The most common workarounds for BI performance involve finding ways to schedule, pre-process or cache reports and dashboards. Eventually the nightly data loading and the increased amount of batch pre-processing for BI can expand beyond the available window. When this happens, no amount of moving work will resolve the difficulty.

Buying more hardware is sometimes the right answer to these problems. More often, adding hardware is a temporary solution because the challenges are with the database, not the

hardware. Several people mentioned that their problem was only at one or two peak times each day, with servers barely busy the rest of the day. A workaround given by several companies was loading data or scheduling queries by time zone to decrease the peak workload. Cases like these are often better solved with different software rather than more hardware.

Scalability

Scalability is consistent performance over growth in elements of the workload, adding an element to simple performance. The classic symptom of scaling problems is deteriorating performance over time. The usual solution is regularly scheduled hardware upgrades. These help when they are first installed, but performance declines as the data grows, leading to an annual cycle of declining performance until the next upgrade.

The most obvious area of growth is data volume, forcing the database to process more data for the same number of queries. Many organizations said that data growth was predictable until the business required new information, for example merging web site operations with the existing transaction data. These additions create big capacity increases outside the normal plans and can slow a well-performing system.

One third of the respondents to the survey referenced in Figure 2 said they have problems related to scalability, equally divided between growth in data volumes and in number of users. User growth and BI maturity are almost as important as data growth. Any increase in use will raise the query load, causing problems even with the same data volume.

One BI manager said, "Success nearly killed our system." He described how they planned for growth in the number of users, but didn't expect to see changes in users' habits. As the users became familiar with the tools and learned what was available, their use increased, exceeding the capacity of the system. As performance worsened, complaints grew, and BI use declined.

Popular new tools for data visualization or exploration are also increasing the load on the database. These tools can be more data-intensive than traditional BI products. Even simple maturity of the BI environment contributes to performance troubles. Reports grow in complexity, and data models become more varied and detailed, creating more work for the database.

Growth of data, users and usage leads to rising hardware and database complexity. Complexity creates a need for deeper expertise and more staff time, adding to the cost of the environment. Eventually most organizations reach a limit of what can be done without radical changes. It's at this point that BI managers look at alternatives to Oracle.

Scalability and cost are related. Adding hardware may not be expensive, but the cost can balloon when you factor in software licenses on the expanded hardware. As one manager said, "Eventually growing the hardware and licenses in our Oracle environment was way too expensive to justify. We were always near capacity. We looked at warehouse-specific databases because they were easier and cheaper to scale."

Cost and Complexity

The cost of performance or scaling is a factor everyone considers, but there is another problem many larger organizations describe: the cost of complexity in IT. A large, varied organization can have several data warehouses supporting different lines of business and many smaller data marts. These may have grown organically, built to support the needs of different departments, or they may have come about through corporate acquisitions.

Data marts can also be viewed as a workaround to the scalability problem. It can be easier and cheaper to give different groups their own data marts, focused on their specific needs, because the incremental cost of expanding and managing a difficult-to-scale central database is too high. As one BI director said, "If you have a data mart with performance problems, then adding to it is unlikely, so it stays a mart, and you add another mart."

The challenge with these environments is not the cost of managing single systems, but the total cost of the environment. While individual systems address specific needs, they introduce redundancy and complexity. Each one extracts data independently of others, causing duplicate ETL maintenance and impacting the source systems more than necessary. They often use different BI and ETL tools, and different databases and hardware vendors.

The secondary costs can be just as high. Multiple systems introduce conflicting information because they extract data from different sources, get it at different times, or apply different rules to clean it. This adds a burden of reconciling data to both IT and users.

Regulatory compliance and security costs related to data rise in distributed environments because there are more components to monitor. The software and personnel costs can outweigh the benefits of smaller databases under departmental control.

Most organizations eventually decide to simplify the BI environment, either by reducing the number of hardware platforms and databases, or else by consolidating the data warehouses. Attempting a large scale rehosting or redesign exposes the scale of operations, and is the driver causing them to look at scalable analytic databases as a replacement platform.

Adding Capabilities

Some organizations have stable data warehouses with good performance and scalability. The change that forced them to look at database alternatives was the need to add new capabilities to the environment. These can add more stringent performance requirements, add significant processing to the database or be incompatible with the existing BI environment.

Several energy companies used the example of new smart grid systems which require real-time loading of data and very large data volumes, processing they couldn't manage on their existing Oracle platforms. Other examples given were new or more frequent data mining, interactive data visualization, operational BI and real-time dashboards.

Any low-latency data requirement, whether for operational BI or a customer self-service portal, requires fast and predictable response times and frequent loading of new data. Mixing this workload on the same Oracle database that supports traditional BI is difficult to manage.

One retailer described how they did market basket analysis once per week on the weekend because to do it on a weekday would slow the BI system enough to make it unusable. After

migrating the warehouse to a Teradata data warehouse, they were able to do market basket analysis nightly, improving their merchandising and promotions. Many companies are now adding data-intensive analytics to their BI environments, driving the need to look at new platforms.

The common element in all of these scenarios is the need for predictable performance. Some require the ability to scale up to meet growing workloads, while others simply need to improve performance to meet existing service level agreements. The next section describes the different approaches organizations took to migrate their data warehouse off Oracle.

Common Approaches to Data Warehouse Migrations

After choosing to migrate BI and analytics work from Oracle, the next big decision is about the approach. There are different approaches based on the problem you're trying to solve and the tradeoffs you want to make. Some take more up-front effort while others require more work after the system has been moved. The following list categorizes the approaches taken by customers who migrated their BI environments from Oracle.

Rehost - Sometimes called the forklift upgrade, this approach moves an existing database to a new platform with no changes except those required to keep the system running.

Consolidate - Several data marts are moved to the new platform. This can be done in two ways: a physical platform consolidation which rehosts and runs the data marts with no change, or a logical consolidation which could involve changes to underlying ETL or shared data models while retaining the autonomy of the data marts.

Redesign - The system is redesigned to take advantage of the new platform's capabilities, which includes changes to ETL, data models and the BI layer.

Augment - Rather than change or move an existing system to add capabilities, the new database is used to run new workloads or offload workloads that don't perform well. This leaves some BI activities on the old system.

Let's look at these approaches in more detail to see how they can be applied to the common scenarios people encounter and what customers experienced when doing the project.

Rehosting a Mart or Warehouse

More than two thirds of customers interviewed approached the project as a platform rehosting effort. The goal was to change the platform to alleviate existing performance and scalability problems, not add new features or support workloads.

The reasons vary for choosing this approach over others. The primary advantage is a short project duration. Minimizing work means avoiding data model changes whenever possible, since they force changes in the ETL and BI layers. When you can simply re-point the ETL and BI servers to the new database, there is minimal change and lower cost since there is less labor.

One company estimated the BI effort to change the data model compared to leaving it intact. They reported that "[It was] a few days of small changes to the BI layer rather than several weeks estimated to change and retest all the reports for a major data model change."

However, the process is not as simple as exporting the data from Oracle and loading it into the new database. Everyone who rehosted said they had to make small changes to their data models. These were required because of the way the Teradata Database distributes data across nodes to enable scalability, and because of small differences in data types.

One of the disadvantages of leaving the Oracle data model unchanged is that it can limit performance and scaling or prevent the use of BI-specific features in the Teradata Database. The original data models were designed to work within the constraints of the Oracle environment, and those constraints are different in a new database.

The size and scope of the existing data warehouse does affect the difficulty of rehosting. The important factors are the number of sources, style of ETL and complexity of the data model. In general, the larger and more complex the warehouse, the less effective a simple rehosting is to be when compared to other migration methods. While changes are minimal, they still add up and may favor a redesign.

The biggest value to a straight rehosting is that the performance gains are immediate. Users see faster queries, or data loads get done well before their deadlines, resolving the visible problem. One BI manager said that a side benefit of simpler scalability was that they spend much less time in meetings with systems and storage administrators arguing about how to tune Oracle and whose idea is right. He said, "My best staff could do more important work than rewriting queries so they'd actually complete."

Most organizations say that the big lift in performance means that losing a little performance not making changes is a worthwhile tradeoff. Aside from resolving the immediate complaints, it reduces the risk of project delays, cost and technical problems.

The caveat to rehosting is that much of the complexity designed into the old environment is brought to the new platform. Most people who rehosted began making changes to the database and applications or started a redesign project within two years. As one architect said, "Rehost versus redesign is a tradeoff. You can pay me now, or pay me later. We didn't recognize this at the time."

Consolidating Business Intelligence Silos

Most people who consolidated multiple data marts onto a Teradata platform were less concerned about performance than they were about the cost and complexity of their BI environment. They wanted the ability to run separate data marts but manage them as a single platform.

The most-reported driver was the value they expected to see by reducing operating costs. In order to achieve this through consolidation, they required a level of scalability that was unavailable in their other databases, leading them to a Teradata solution.

There was a divide in how people approached a consolidation project. Some said they treated the consolidation as a set of rehosting projects, moving data marts to the new platform with little to no change. There was sufficient value to justify the project by reducing the number of servers, storage, licenses for the database, ETL and BI, and the associated operational costs.

Others had a different view. They felt that the bigger cost driver was the complexity of data integration and data management in the siloed environment. This meant they spent time identifying shared data and combining ETL into a repository that fed the marts. They designed a common staging and management layer from which to populate shared data, leaving the data models of the marts unchanged so their users would be unaffected.

Several companies said that the rehosting model for consolidation saved them money and time, but it was less valuable than they expected. After the one-time savings from server, license and system management reductions, they still had to invest too much in labor for integration.

The difference in view can be partly explained by the variation in organizations. Many of the organizations that treated consolidation as rehosting had more local autonomy or diversity of BI and data needs. The organizations that changed the data management layer had more common data and usage.

One bank described the process it went through: "We had multiple data warehouse environments, some on Oracle, some on SQLserver, one in DB2. We did a forklift move of environments to Teradata as fast as we possibly could. We didn't get the results we wanted because we were treating it as a co-located set of independent marts. Also, these had been optimized for Oracle right down to the staging areas for ETL, giving us a performance penalty that reduced the overall capacity we expected from the box."

The message from these companies is that mart consolidation works and saves costs, but it's less valuable when treated as a simple platform rehosting. A unified operational data store on a Teradata platform can simplify ETL and data movement to marts hosted on the same platform, even if they have their own target schemas and BI. The key elements are whether they have different sources and whether their data needs to be viewed in the aggregate. If so, an independent pool of marts may be the better solution.

One question that organizations needed to answer when starting their consolidation was how to determine which marts to migrate and when. Some chose to treat the effort as one "big bang" consolidation where they would move all the marts as quickly as possible. Others opted to migrate incrementally over a longer period.

Key factors were the size, complexity and number of marts. A high degree of any one of these favored incremental moves. The biggest driver for the "big bang" approach was trying to take advantage of costs savings in existing hardware upgrades and license renewals that had specific deadlines.

There can be challenges because some groups don't want to give up the control they have over their data mart. The best advice on consolidation was from one IT manager who said, "You can't be militant about marts. You have to be pragmatic. The best thing to do is look at their hardware upgrade cycles, or see when they have a big change coming. Often that disruption is what opens the door to them wanting to work with you on the shared platform."

Redesigning the Data Warehouse

One of the triggers that causes organizations to explore new databases is redesigning their data warehouse as an attempt to meet performance goals. A redesign implies ETL and BI changes too, so it's not a big step to consider moving to a more scalable database.

A BI manager at one organization described how an Oracle redesign led him to migrate the data warehouse. "We looked at spending a lot of money on hardware, Oracle licenses and changes to our designs with no guarantee that it will fix the problem. Even if it does, we know we'll encounter the same problem again down the road. This made it easy to justify a proof of concept since we can see the performance we'll get, and we know the cost of scaling up."

More organizations still choose to rehost rather than redesign because a redesign is more work and takes longer. They often redesign incrementally after the initial rehosting, an evolutionary approach to migration. Those who do are often looking at more than just poor performance.

Sometimes redesign is driven by evolution of BI in the organization. As use matures, needs become more diverse and the data more detailed and varied. Other times, it's a change driven by new practices, like integration of web analytics data into the existing warehouse model.

Reducing complexity and cost are common goals as part of a redesign effort. The experience of the following electronics manufacturer is representative of the evolution many organizations have seen in their environments.

The company started with one Oracle data warehouse and later added two more to support overseas operations. These were added because they couldn't load data nightly for one region without ruining performance for the users just starting their workday in another region.

This went well initially. The systems were exact copies of each other running on different hardware. Providing a single picture of corporate operations was a challenge because the data was spread across three systems. All the regions developed downstream systems over time, increasing the complexity. Growth in data and users in one region caused them to change their data model, while another region replaced source systems and changed their ETL.

The systems diverged enough that reporting across all three was a problem. Both business and IT staff had to spend hours each week reconciling data for corporate reporting. For this they created another reporting application to reconcile the three systems. Adding to the cost were performance problems that took advanced database and system administrator skills to resolve.

Eventually the CIO decided that the BI environment was out of control, and they needed to do something different. "Our per-user cost of BI was rising across the board due to the complexity. I wanted to change that so we could scale more affordably in smaller increments." His vision was to provide a consolidated information platform to manage the extraction and delivery of data and to consolidate reporting where possible in an enterprise data warehouse.

They did a complete redesign, consolidating the regional systems back into a single Teradata Database. Costs are down, and according to the DBA manager, "Administration time took one third of what was needed on the Oracle side."

Redesigning an existing warehouse is not easy. Whether it's driven by the need to consolidate applications or simply to improve performance, a platform change can be disruptive. A firm that redesigned and consolidated to gain control over sprawling silos discovered they had so many downstream systems running against one of the marts that they had to run it in parallel for a year to migrate all the data feeds. Their advice: "Watch who connects to the database before trying to rehost or redesign a mart."

Augmenting the Existing Data Warehouse

Sometimes the performance or scalability of the Oracle environment meets the needs of the existing BI workload. It's only when a new application or type of processing is added that performance problems appear. Rather than migrate the data warehouse, organizations choose to run new workloads on a new platform and leave the legacy BI environment in place. In some cases, they later migrate the BI to the new platform after it has been proven to work well with the new workload.

The desire to add advanced BI capabilities is one common reason cited for augmenting the existing Oracle environment. This includes work like data mining for scoring or segmenting customers, calculating product affinities, or analyzing text. In many cases, the data-intensive processing is greater than the current system can handle so it's offloaded.

Data mining isn't the only reason to augment the environment. Another common need is for low latency BI, for example in a call center or to support an online customer self-service application.

With operational BI and other low-latency processing, one reason for offloading is that data is being continuously loaded while being queried. It's easier to design a system on a Teradata system specifically for this than it is to mix real-time data loading, large numbers of small concurrent queries, and a traditional BI workload on an Oracle system.

This type of workload imposes other needs. Customer-facing operational systems like these have strict service level agreements. It can be less expensive to maintain a separate platform with the required resources in a high-availability configuration than it would be to upgrade an entire data warehouse.

A communications company explained their rationale. They wanted to run a specific set of analytic models on customer and call data. When they tried to do this in Oracle, it would take one week to process a few months of data. It was impossible to score the entire customer base.

They tested the same processes on the Teradata solution and were able to reduce the time to a few minutes. They chose not to migrate the existing data warehouse because the project cost would have been prohibitive and because they were worried about mixing the heavy batch processing load with interactive BI. Instead they built a separate application on an appliance to run that workload.

When the need is specific and isolated, offloading in this fashion works well. If the desire is to add broader capabilities, then the approach may not work as well. In this case, the analytic capabilities should coexist with existing BI on the warehouse platform. Separating the workloads can add significantly to operational costs because data must be moved out of the data warehouse to the new system doubling the data footprint and nightly ETL work.

The other challenge is re-integrating the results with the BI environment. When on separate platforms, they are reported separately. This means they can't be easily combined with other information. A single platform can manage the workloads involved so data doesn't need to be copied onto other servers, and it allows insights to be retained with the rest of the organization's data.

Conclusion

The root cause of the problems that drive most scenarios is performance or scalability of the existing database platform. These problems lead people to workarounds that have lessening impact and result in more cost and complexity over time while delivering less value.

The root cause of these problems must be addressed. The database is the heart of the system. Avoid the workarounds and focus on this rather than spending time and money on repeated tuning, hardware upgrades, BI and ETL changes. These workarounds lead to increased labor and hardware costs and missed business opportunities. You want to invest in technology that will position you to be at the low end of a highly scalable platform rather than pushing you to the high end of a less scalable platform.

The companies interviewed had some common recommendations for planning and managing migration projects:

- The best time to evaluate a new platform is when there is a discontinuity in the IT environment like a business reorganization or the need to do a major hardware refresh. Any time there is a disruption to the environment, whether due to changing business needs or technology, it will be more acceptable to look at alternative database approaches.

- A hard lesson for many was the cost of deferring problems caused by complexity and silos of information. The companies who faced this all agreed: start earlier so that you'll be able to address the situation sooner rather than dealing with a bigger problem later. By that time, the pressure to do something will be immense, and the time to deliver short, causing compromises along the way and increasing the duration and challenges of the project.

- The impact of a platform change can be larger than expected in organizations where the data warehouse has been in place for a long time. As the environment matures, new systems evolve and depend on the data warehouse. Sometimes these are ad hoc, unsanctioned by the IT department and invisible to the warehouse team.

- If the dependent systems aren't identified early, they can add a lot of time to the migration effort. In some cases, they can require months or even years of keeping the old warehouse on life support while their data and integration requirements are addressed.

This can bring political complications to the forefront. Just because someone isn't supposed to do something doesn't mean they won't do it. Shadow IT groups exist for many reasons beyond the inability to serve their needs. Sometimes the reason is related to control, security or simple interpersonal troubles between managers.

These will surface when the migration affects downstream applications, or when the project is a consolidation of multiple marts or warehouses. These changes affect the roles of people in other departments, change budgets and organizational structures. It's important to understand these impacts in advance so the changes are not a surprise to the people affected.

All interviewees felt that the move away from a general-purpose database to a Teradata Database was worth the effort. Platforms designed to support analytic workloads avoid the performance and scalability problems traditional databases have with these workloads. Analytic databases provide a way to support existing BI needs while creating new possibilities for use of data.

About the Author

Mark Madsen is a research analyst focused on analytics, business intelligence and information management. Mark is an award-winning architect and former CTO whose work has been featured in numerous industry publications. He is an international speaker and author. For more information, or to contact Mark, visit http://ThirdNature.net.

Third Nature is a research and consulting firm focused on new practices and emerging technology for business intelligence, analytics and information management. The goal of the company is to help organizations learn how to take advantage of new information-driven management practices and applications. We offer consulting, education and research services to support business and IT organizations and technology vendors.

Sponsored by

Third Nature, Inc.
PO Box 1166
Rogue River, OR 97537

EB-6437 > 0911