The Challenges of Implementing a Data Warehouse to Achieve Business Agility

U.S. Symposium/ITxpo 2001

Kevin Strange

7–10 May 2001
Colorado Convention Center
Denver, Colorado

56 Top Gallant Road
P. O. Box 10212
Stamford, CT 06904-2212

Conclusions

- **Preparing data for a data warehouse is complex and requires resources, strategy, specialized skills and technologies.**
- **The ETT tool market is undergoing significant change and turmoil, making most tool investments tactical.**
- **Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure.**

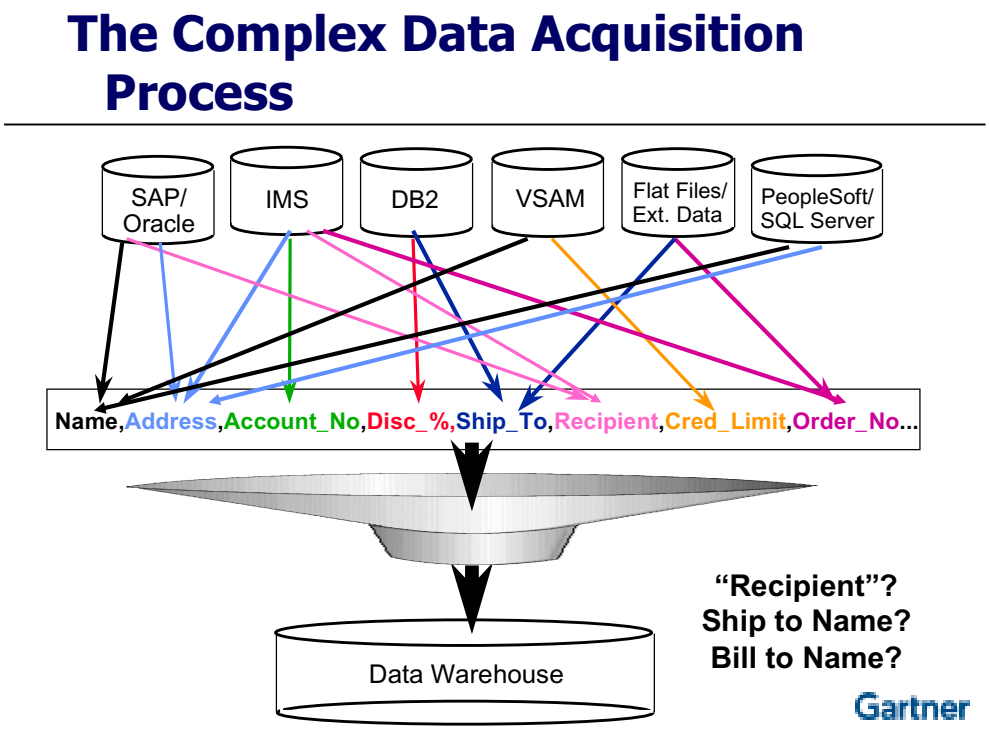
Source: Gartner Research

Implementing a data warehouse (DW) is a daunting challenge — one that is often underestimated and requires enormous amounts of time and resources. Creating a multisubject, consolidated information store requires reconciling the different data models that run the organization's highly heterogeneous operational database management system (DBMS) environment — data models that have been individually designed and implemented in the last 20 to 30 years. Furthermore, the operational data must be standardized, integrated, enriched and made consistent. In addition, any time data is moved, data inconsistency and integrity exposures are created.

DBMS decisions are often made with simplicity in mind, and many organizations default to their choice of DBMS used for online transaction processing (OLTP) applications. Traveling this route for technology selection will be filled with potholes and wrong turns, greatly increasing the risk of failure.

A number of technology decisions need to be made and the right decision can reduce the challenges, which can dramatically increase the chance of success. In this presentation, we will explore strategies for using these technologies successfully to reduce the magnitude of risk.

Tenets/Principles

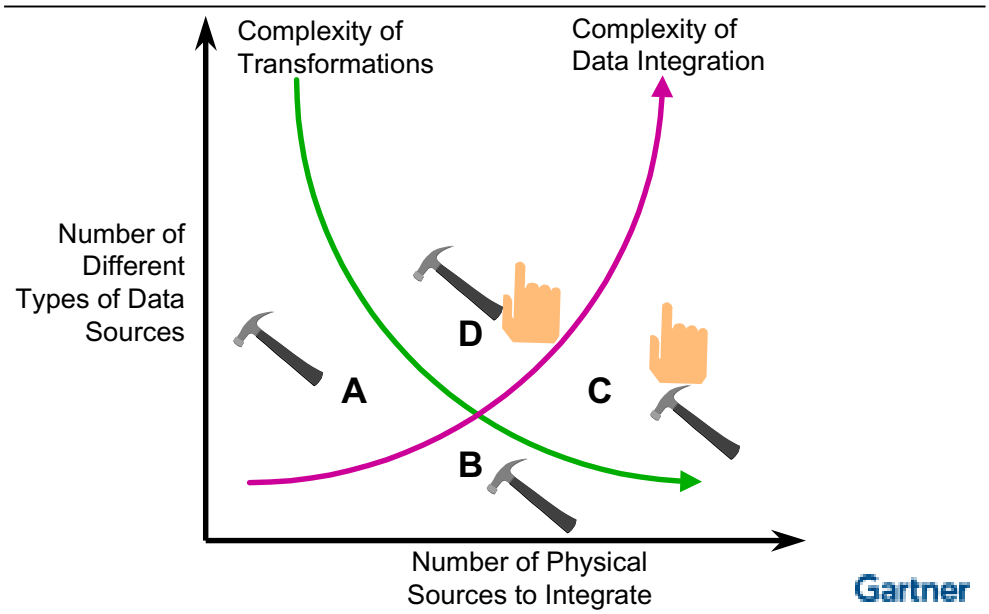


Source: Gartner Research

Conclusion: Preparing data for a data warehouse is complex and requires resources, strategy, specialized skills and technologies. Many factors contribute to the complexity of preparing data for the DW. Most organizations have an average of eight different operational DBMSs, with easily more than 50 databases from which data will be acquired. Accessing these heterogeneous databases requires special skills and technology to deal with the different syntax of the DBMS. Different semantics also add to the complexity. Understanding the operational data model and the meaning of its data elements is an intense analytical effort. Typically, the definition of data elements, e.g., "customer," is not consistently implemented across the multiple, stovepiped operational databases. Furthermore, for older operational sources, organizations no longer have personnel skilled in legacy systems and databases, nor do they have much documentation describing the underlying record structures and semantics. The steps for preparing data for the DW are: *1) Data identification:* Identifying the most appropriate operational source(s) from which data element(s) can be acquired and analyzing the formats, potential duplication, potential incremental value and source reliability. *2) Data acquisition:* The physical collection of the data. *3) Data cleansing:* Processing designed to eliminate duplicates, correct erroneous values and replace missing values, and any other transformation to improve the quality of the data. This is usually done prior to loading to reduce load failures. *4) Data transformation:* The process of enriching data into more-business-descriptive content. 5) *Data integration:* The need to focus on subsets of data based on subject areas, not sources, needs to be undertaken, which will require complex processes that integrate data from multiple data sources.

Decision Framework

By Hand or With a Tool?



Source: Gartner Research

Conclusion: Preparing data for a data warehouse is complex and requires resources, strategy, specialized skills and technologies.

The total complexity of the extraction, transformation and integration process depends on the number and variety of data sources. We propose the mathematical product of the number of sources of each type as a first-order approximation of complexity. If this is roughly accurate, then the more data source types that are introduced, the more complex implementing the resulting data acquisition process becomes. By integrating data from legacy applications with newly deployed client/server applications, users risk adding a level of complexity that will make the extraction and integration task nearly impossible to perform via traditional application methods without a large amount of resources and a complex, multistep process that will be difficult to maintain (Point C and Point D above). Enterprises facing such challenges must plan to manually augment the functions automated with any of the available extraction/transformation/transport (ETT) tools. Enterprises will continue to expend resources because the complexity of the data integration problem cannot be completely solved with off-the-shelf technology.

Action Item: Enterprises must maintain programming skills for pre-relational DBMSs if source data will be required for extraction and manipulation for the DW.

Imperative: As the number of data sources and types increases, the complexity of the extraction and integration process increases, and requires the appropriate amount of resources to implement. Application support teams must be involved to "protect" the data center's resources and ensure quality data for the data warehouse.

Resource Demands: Extraction and Integration

Source: Gartner Research

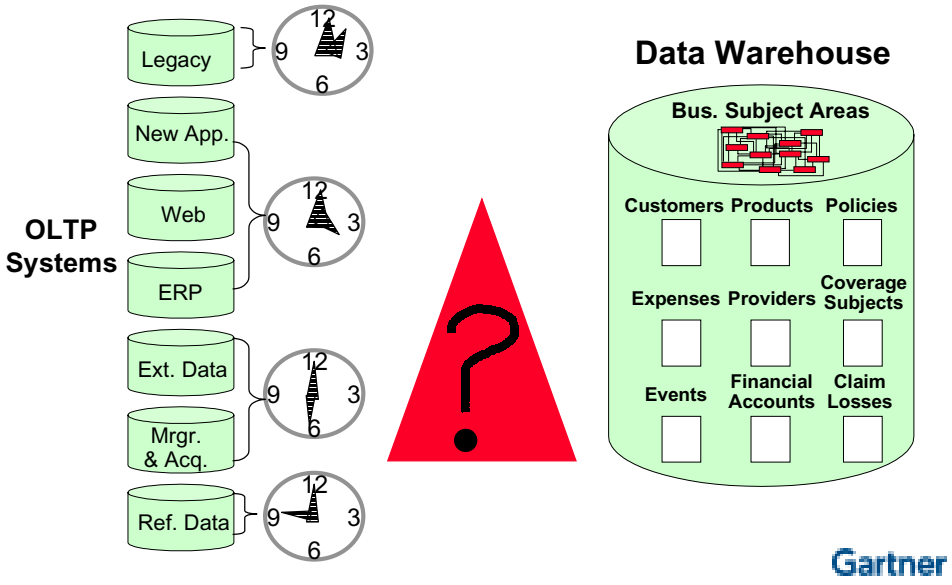
Conclusion: Preparing data for a data warehouse is complex and requires resources, strategy, specialized skills and technologies.

User often try to solve the extraction, transformation and integration problem with additional resources, in addition to those that already exist and could be easily leveraged. Looking for magic instead of knowledge, enterprises miss the point of the issues involved in performing this task. That is, users will continue to expend resources even though the problem cannot be completely solved with off-the-shelf technology. Vendors providing extraction and transformation tools must prove the viability of their solutions at the scale of their target marketplaces. In addition, enterprises need to have the support of application teams to provide not only the first stage of the data acquisition, but also the agreed-on files to the DW technical staff, which will incorporate the process to perform further integration, transformation and cleansing to ensure the accuracy and quality of the data in the DW. In addition, the data center must provide support for this process to ensure the scheduling and timeliness of the data movement to the DW. In a "multicolored" data center, this will be a significant challenge.

Action Item: Management commitment must be given to allocate the required time of application support teams to the data preparation phase of the DW implementation project.

Strategic Planning Assumptions: Through 2005, the time boundary for refreshing the data warehouse will remain a nightly batch process (0.8 probability). By 2005, 80 percent of an enterprise's transactional applications will share an integrated, consolidated database, making near-real-time updating of the data warehouse a reality (0.2 probability).

How Current Is the Data?



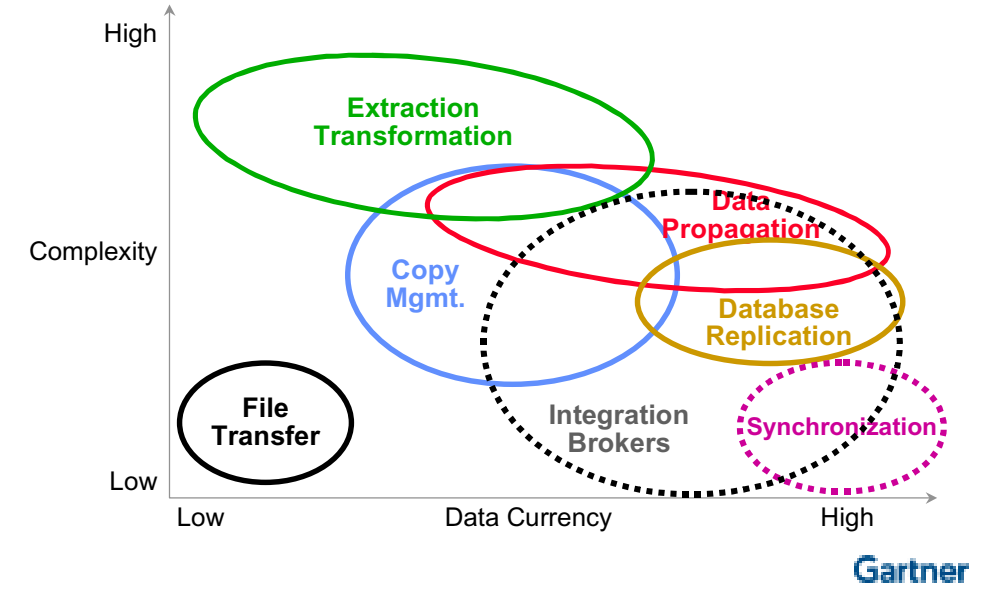
Source: Gartner Research

Conclusion: Preparing data for a data warehouse is complex and requires resources, strategy, specialized skills and technologies.

A DW includes all of the data for all of the subjects that are of shared interest to business users. Data from a variety of transactional systems must be integrated, reconciled and normalized for loading into the DW's subject-oriented data model. Since all of the transactional systems operate on different schedules for updating (i.e., some are online, have 24x7 availability and are continuously updated, whereas others have a nightly batch update cycle), often crossing different time zones around the world, the DW team must pick a somewhat arbitrary time boundary for updating the DW. This time boundary has moved from monthly to weekly to nightly. Because a DW is designed to support strategic decision-making (distinguishing it from decision support databases designed exclusively to support operational, day-to-day decision-making), data should be consistent and not constantly changing. Strategic analysis usually requires some historical data and the capability to identify trends and patterns. If the data is constantly changing, this makes comparative analysis or strategic planning and modeling types of analysis impossible. Therefore, we do not expect the time boundary to get any more current than the close of business yesterday. An update strategy must reflect the frequency and the granularity of updates (adds, deletes or inserts) that are desired. This will determine the methodology and the technology choices for updating. *Action Item: Enterprises should carefully assess business users' requirements for "timely" information; up-to-the-minute information is rarely needed.*

Decision Framework

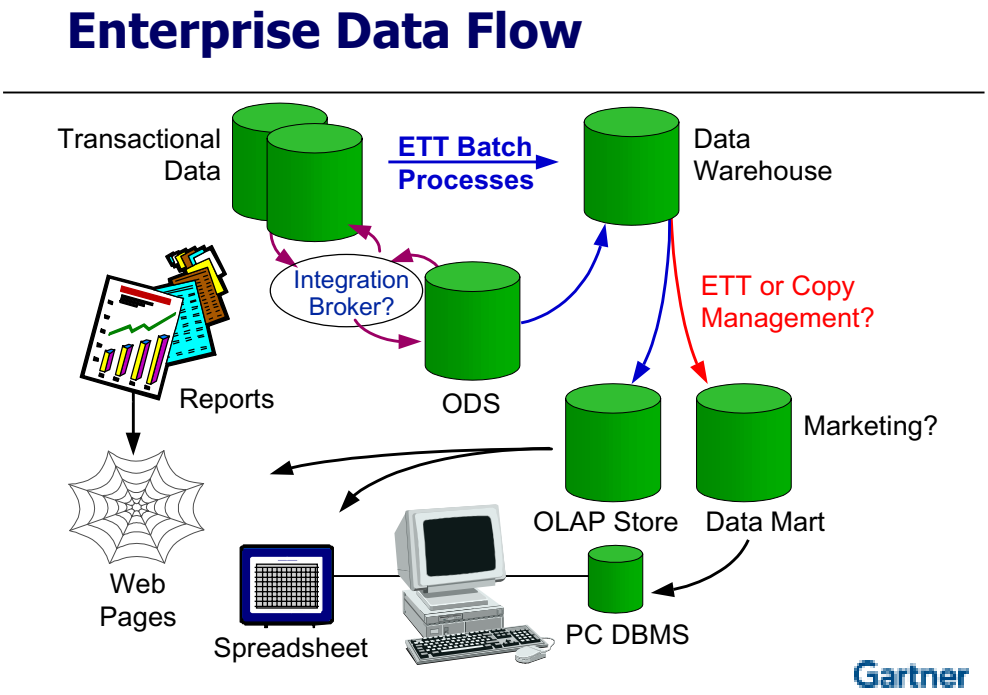
Data Latency Perspective



Source: Gartner Research

Conclusion: The ETT tool market is undergoing significant change and turmoil, making most tool investments tactical. We position alternative data-movement technologies along two dimensions: complexity and data currency. *Complexity* encompasses the number of sources, the degree of heterogeneity and the degree of transformation. Organizations must make a trade-off between these two dimensions whenever data is moved; the more data is transformed, the longer it takes and, thus, the higher the latency. *Data propagators* propagate data between heterogeneous DBMSs or schemas. *Replication* is between homogeneous DBMSs and schemas. Data *transformation* tools change data syntax as well as the contents (the values) and even the semantics (the meaning of the data). They apply logical algorithms, using a programming language other than a data manipulation language (e.g., SQL), to transform data for use in another application environment. Data *synchronization* is an application profile, not a technology. True synchronization requires a two-phase commit protocol. Replication reconciles updates between two or more databases, albeit for a fleeting moment, as users continue to make changes to their respective databases. Replication and message queuing can be used for application-specific synchronization. Regardless of the technology, business rules must be programmed to drive the reconciliation and are critical to the application's success. *Copy management* is the same technology as replication. The difference is in the implementation; it is a scheduled snapshot in time, thus resulting in a consistent result set. It is appropriate when the order of transactions is less important and lower overhead on the source database is desired.

New Rules/New Realities



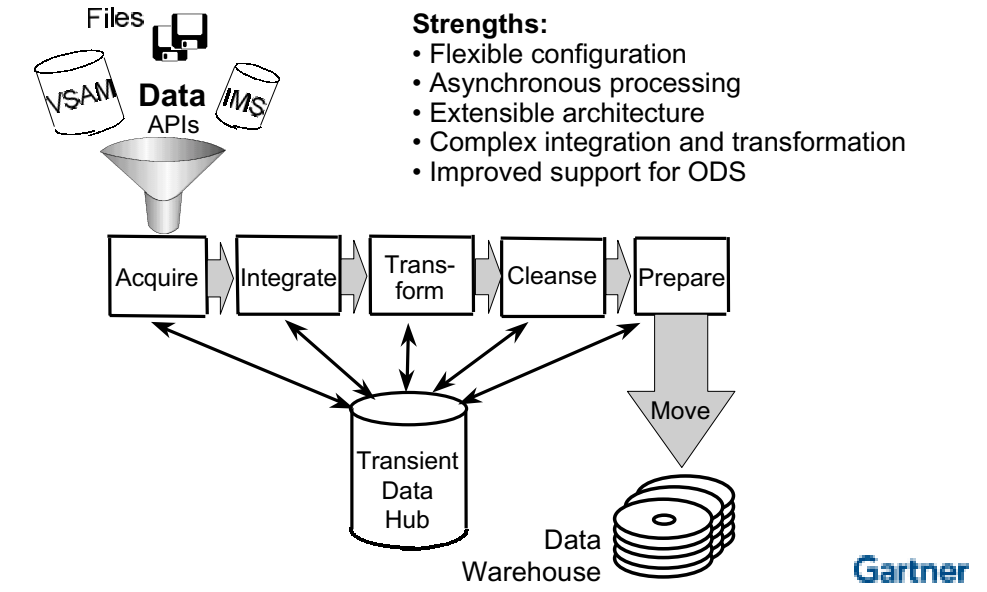
Source: Gartner Research

Conclusion: The ETT tool market is undergoing significant change and turmoil, making most tool investments tactical.

Modern data architectures are looking more like oil refineries, where data is moved from one specialized server to the next. As in the oil refinery, emphasis should be on added-value processes, or the conversion of data to a new structure, so that data is not merely moved for the sake of moving it. Although centralized, shared logical data models are recommended foremost, they are often impossible for reasons having nothing to do with sound data-management principles. Thus, successful organizations will determine the right compromises between where data movement adds value (e.g., performance, simplicity and interfaces) and where it subtracts (e.g., security, accuracy, shareability and auditability). Understanding the qualitative changes that occur when data is moved or transformed is essential to good data-architecture planning. Data architectures like operational data stores (ODSs) and DWs not only consolidate data that originates in stovepiped applications, but also data is semantically enriched along the way to these repositories, increasing its business value. As data moves within the refinery, it is continually transformed. Success of the refinery, as measured by enterprisewide information sharing, requires metadata management, data lineage tracking, audit trails, security policies, data quality procedures and impact analysis functionality — all features lacking in most integration middleware technologies today. *Action Item: Organizations should assess whether the value of refining data beyond the ODS or DW outweighs the exposures created by further decentralizing the data.*

Strategic Planning Assumptions: By 2004, transformation engine technology will be used in 75 percent of all application integration designs, due to its strong capability to reconcile the semantics of data elements using declarative business rules (0.7 probability). Through 2004, transformation engine technology and integration broker technology will remain completely separate, without any degree of integration (0.3 probability).

Transformation Engine Architecture

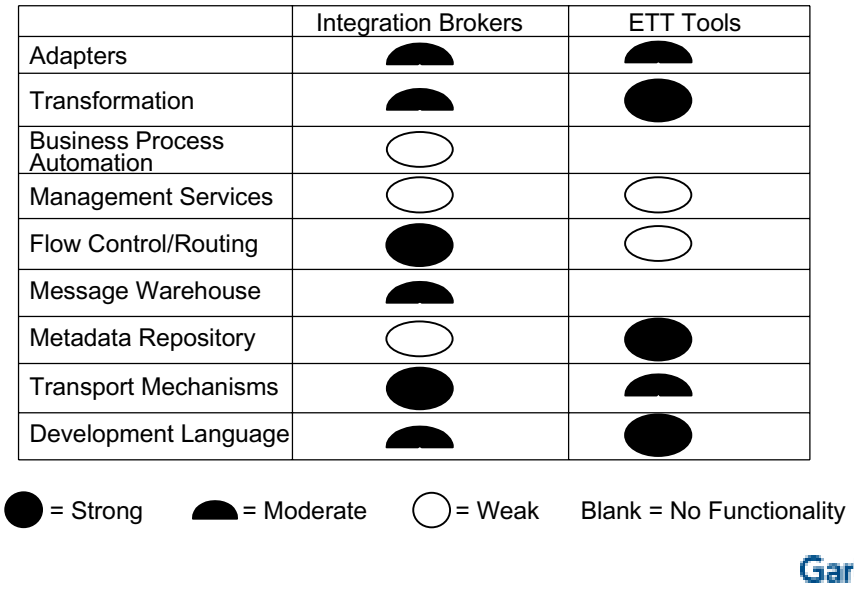


Source: Gartner Research

Conclusion: The ETT tool market is undergoing significant change and turmoil, making most tool investments tactical. In 1996, we defined the transformation engine (TE) architecture. TEs provide a multitasking architecture to perform multiple functions (acquire, transform, cleanse and load) asynchronously and in parallel — functions previously performed serially with batch application programs. Each function performs at scheduled time or event intervals. With a runtime engine that can be deployed on separate processors from the production OLTP systems and the target DW or other decision support system (DSS), the TE can leverage the scalability of symmetric multiprocessing (SMP) hardware and provide greater performance and throughput than batch applications. This architecture can provide an improved ability over previous batch applications to handle a large number of data sources and highly complex transformations. Its strength is its multitasking design, which uses a data hub as a transient store or staging area. However, most of the TE-based ETT tools on the market today have little support for nonrelational data sources. They only access relational DBMSs (RDBMSs), flat files or Open-Database-Connectivity- (ODBC) accessible databases; other sources must be manually unloaded or accessed via a gateway. Nevertheless, the rate of adoption of TEs compared to code-generating ETT tools is approximately 10-to-1. This architecture is also used in most integration brokers, which capture individual messages or transactions at the application level, via an application programming interface (API), rather than directly from the database. *Action Item: TE-styled integration brokers should be considered for meeting application integration requirements.*

Strategic Planning Assumptions: By 2004, 80 percent of "Fortune 5000" enterprises will use at least two transformation-engine-based application integration tools — one to meet their transactional interfacing requirements, and one for their batch data integration requirements (0.8 probability). By 2004, transformation-engine-based products that are equally good at near-real-time and batch application integration will be in widespread use by the "Fortune 5000" (0.2 probability).

Integration Brokers vs. ETT Tools



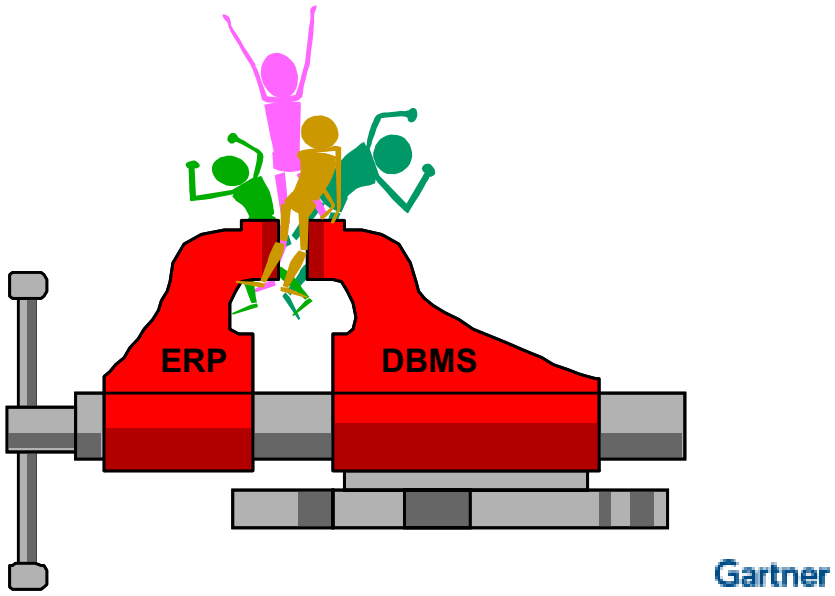
Source: Gartner Research

Conclusion: The ETT tool market is undergoing significant change and turmoil, making most tool investments tactical.

ETT tools and integration brokers have much in common, but they also have important differences. ETT tools use DBMS protocols to talk database-to-database, whereas integration broker products communicate application-to-application (and sometimes application-to-database) using application-level APIs. Integration brokers capture individual transactions or messages (a verb plus data) and route them to a receiving API. They capture updates in the context of transactions or messages rather than as a set of database updates. In this way, the transactional context and transactional integrity of the receiving application are preserved (since new updates are subjected to the same validation and editing rules as native updates.) These tools integrate individual or batch transactions. In contrast, ETT tools always integrate sets of data (collections); they circumvent application-level logic by directly inserting SQL updates at the database level. Thus, data integration designs are only appropriate for "loosely coupled" applications. A DSS and its related OLTP counterpart(s), or a master file and OLTP applications that reference it, are examples of "loosely coupled" application systems. Both categories of tools have optimized their products for one of these two interfacing designs. *Action Item: Enterprises should avoid the temptation to apply one integration product to the other's problem domain, unless the tool is leveraged across multiple projects. Doing so will require significant customization and extensions to supplement the tool's off-the-shelf functionality.*

New Rules/New Realities: Two or three independent transformation-technology providers will effectively compete with DBMS vendors by providing advanced and specialized features in their transformation tools, such as heterogeneous DBMS support, packaged-application API support and unique transformations for specific applications.

ETT Vendors Get Squeezed Out



Source: Gartner Research

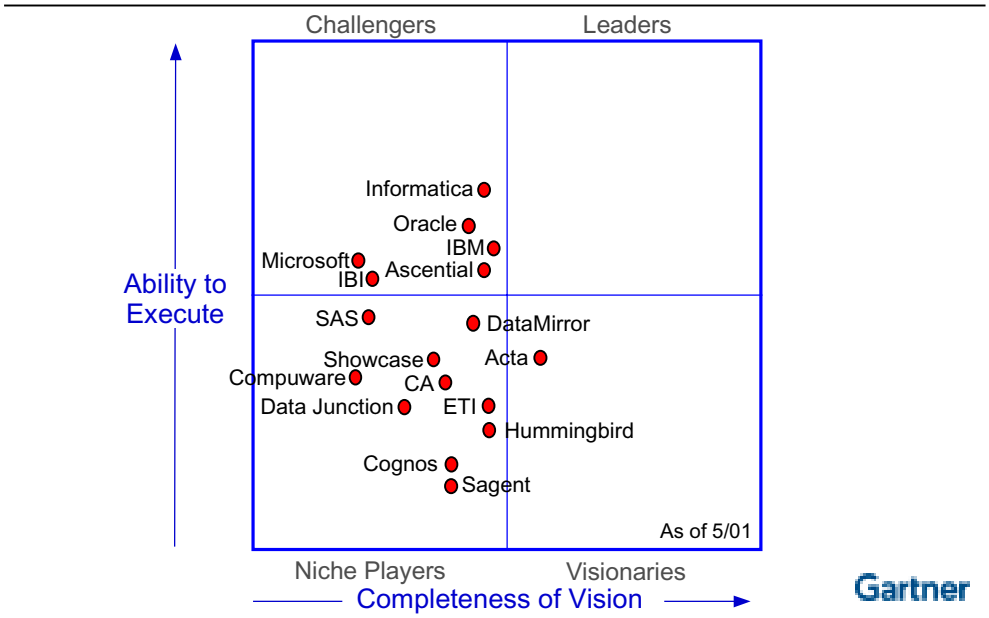
Conclusion: The ETT tool market is undergoing significant change and turmoil, making most tool investments tactical.

Increasingly, enterprises are looking to their preferred RDBMS or enterprise resource planning (ERP) vendor for most of their DW infrastructure. IBM, Oracle, Microsoft, SAP and PeopleSoft claim to offer a complete solution for data warehousing and business intelligence (BI), including ETT functionality. As these vendors focus on dominating these markets, independent ETT vendors are getting squeezed. Many vendors have either been acquired (eight since 1999) or begun moving into adjacent markets. Since the revenue growth potential of this stand-alone market is small, only a few third-party providers will survive this market shakeout and continue to grow. The vendors most likely to remain focused on data integration middleware, including ETT functionality, are Information Builders (IBI) and DataMirror. All the other vendors in this market today are already evolving their products into other markets, such as the BI market segments, packaged DW data models, or the application integration market segments.

Action Item: Organizations should view any infrastructure-independent ETT tool purchase as tactical; a 24-month or less return on investment (ROI) should be realized.

Market

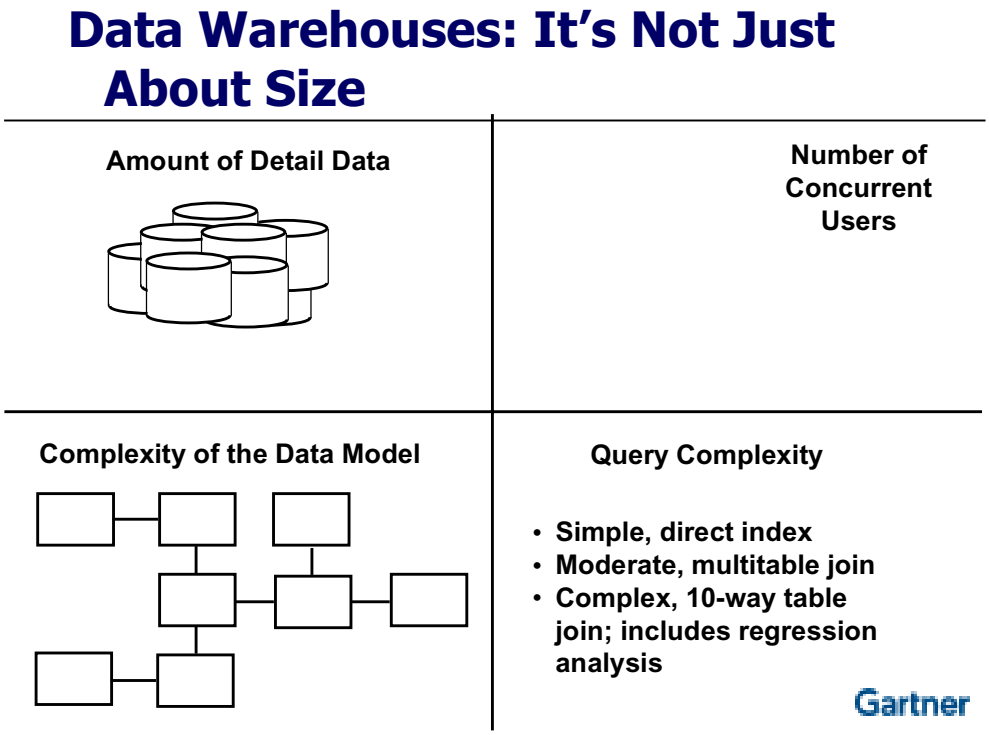
The ETT Magic Quadrant



Source: Gartner Research

Conclusion: The ETT tool market is undergoing significant change and turmoil, making most tool investments tactical. ETT functionality from the leading RDBMS vendors continues to improve and gain market visibility, although somewhat slower than expected. Therefore, there is still a significant gap to be closed before it is considered "as good" as offerings from the pure ETT technology vendors. As the DBMS vendors continue to close this gap, it will be difficult for buyers to justify a premium for the incremental functionality provided by third parties. Third parties will (*must*) differentiate on ease-of-use, better administration and higher performance. They will target the 20 percent to 30 percent of buyers who truly need their advanced functionality and are willing to pay approximately 30 percent more. However, as ETT functionality gets embedded into the DBMS or BI solutions, this market will collapse; it will no longer exist as a distinguishable market with multiple players. This collapse will occur gradually in the next 12-to-24 months before leaders emerge. Therefore, any purchase should be viewed as tactical, with ROI in 18-to-24 months. This still leaves the organization exposed, since applications will be deployed using this technology and have a longer useful life than 18-to-24 months. To minimize this exposure, organizations should choose a vendor: 1) with a strong partnership with the chosen RDBMS, BI, ERP or middleware vendor; 2) that is committed to the strategic development platform; or 3) that appears to be an acquisition target for one of the organization's strategic technology partners. *Action Item: Users should think strategically, even when buying tactically.*

Definition: Although data warehouse DBMS vendors seem preoccupied with size, users need to evaluate the selected DBMS on the basis of four major criteria (i.e., amount of detail data, number of concurrent users, complexity of data model and complexity of queries) to determine its capability to support a data warehouse.



Source: Gartner Research

Conclusion: Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure.

The perception is that the single most important aspect of a DW is its size. All too often, the number of terabytes is used as the benchmark of a DW's effectiveness. While important, examining size alone is too simplistic, and a DW's value extends beyond the volume of data that it can store — even if we can "normalize" the quantification of size. Several other important challenges are often overlooked.

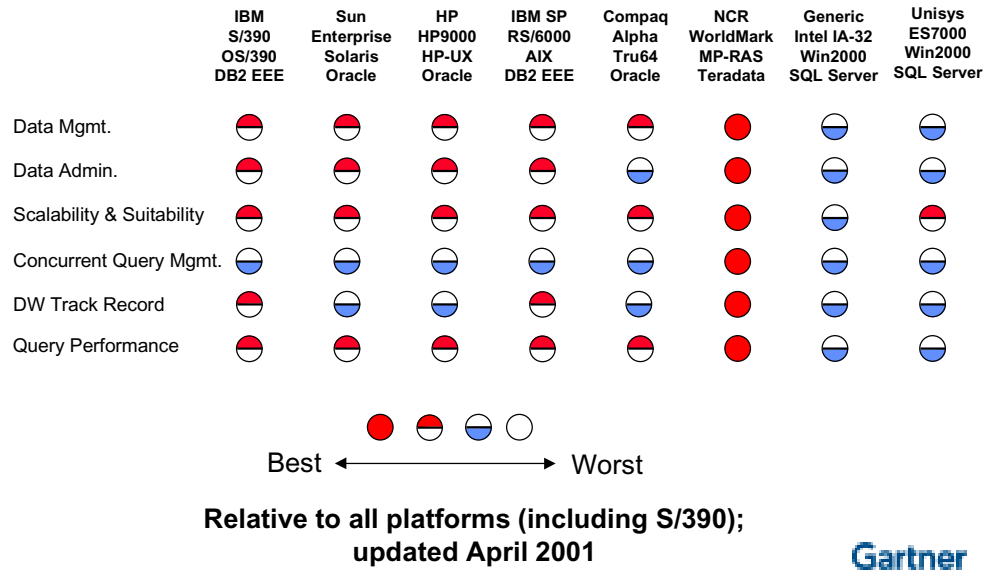
Schema Complexity: Unlike a data mart, a DW focuses on the data, not a specific application. The data model's complexity plays an important role in application neutrality but can make it more difficult to optimize and perform queries effectively.

Concurrent Users: A DW, no matter how big it is, will not be successful if it can only be accessed by one user at a time. The capability to support multiple, concurrent queries in a scalable fashion requires sophisticated workload-balancing capabilities that go beyond simply allowing queries to be "submitted" to the DBMS concurrently.

Query Complexity: Initial use of data by users is fairly simple and involves few tables; but data analysis is heuristic in nature and progressively becomes more complex, and some power users will soon be placing major demands on the system.

Market

ASEM — Data Warehouse DBMS



Source: Gartner Research

Conclusion: Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure. DBMS evaluation criteria are as follows: *Data management:* Does the RDBMS support management of large quantities of data, demonstrating experience in responding to real user demands while maintaining high availability? *Data administration:* Does it feature facilities to understand, predict and optimize resource usage, or provide tools and information for the database administrator (DBA) to manually undertake this task? Are there optimizer facilities? Can the RDBMS cache result sets for commonly run queries and monitor the requirement for summary tables? *Scalability and suitability:* Can the RDBMS provide the necessary flexibility of hardware platform choice —in the range of both available hardware vendors and supported hardware architectures — to provide adequate growth potential? *Concurrent query management:* How well does the RDBMS support multiple users performing a mixture of both simple and complex requests? Does the DBMS provide workload partitioning and balancing? Does the DBMS feature a concurrency model that can support data loads to the database? *Proven DW track record:* There is no substitution for real-world implementations and proof. Can the vendor offer reliable references, preferably from the same industry and with a similar DW profile? *Query performance:* How proven is query performance, such as SQL optimizer strength, ability to join processing algorithms, techniques for handling specialized schema types, indexing methods and data partitioning techniques?

Strategic Planning Assumptions: Through 2004, IBM will continue to be a data warehouse RDBMS technology leader, but its capability to gain market share will be restricted by its lack of credibility on non-IBM server platforms (0.7 probability). By 2004, IBM (along with Oracle) will be a leader in data warehouse implementations on non-IBM server (Unix and NT) platforms (0.3 probability).

IBM DB2 UDB for Unix

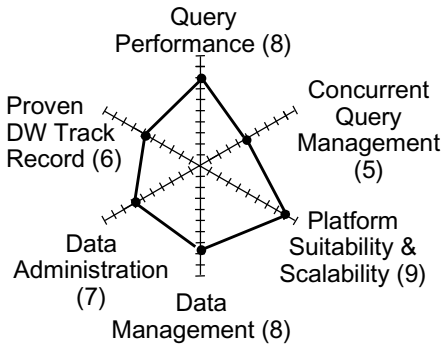
Strengths

- **Performs well on both SMP and DMP systems**
- **Strong query optimization**
- **Leader on NT clusters**
- **Implementation of large data marts and data warehouses**
- **Rich feature set and capabilities**

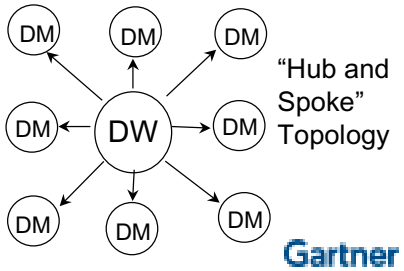
Challenges

- **Lack of non-IBM server platform credibility**
- **Manageability**
- **Unproven with increasing size (very large), complexity and concurrent queries**
- **Rapid release of major versions and product immaturity**

DB2 UDB EEE on RS/6000 SP



Common Topology



Source: Gartner Research

Conclusion: Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure.

Since 1997, beginning with the release of DB2 Universal Database (UDB) v.5.1, IBM has kept the pressure on competition in the DW DBMS market with a succession of releases that included v.5.2, v.6.1 and v.7.1 in each year following. These releases added to the architectural strength of DB2 UDB by providing a plethora of features and capabilities that were geared toward supporting DWs and complex query processing. DB2 UDB is rated a technology leader on the strength of its query technology and the degree to which DB2 UDB Enterprise and Extended Enterprise editions can scale on both SMP and distributed memory parallel processors (DMPP) architectures (Unix or NT). However, the rapid release may have helped IBM to catch up and surpass some vendors on the feature and functionality checklist, but left the utilization and maturity of the new features behind in its wave of releases. Most organizations were not prepared to upgrade their DW implementations as quickly as the releases were becoming available. Thus, mostly new customers that are not in position to implement these features and capabilities in significant production environments implement the new DB2 releases. In addition, IBM still has much work to do in the area of multiquery concurrency to provide support for varying applications and query workloads.

Strategic Planning Assumptions: Through 2005, Oracle will continue to be the market share and mind share leader in data mart and data warehouse implementations, with the support of virtually all third-party application and tool suppliers (0.6 probability). By 2005, Oracle's lack of a comprehensive focus on capabilities to improve support for very-large and complex data-warehouse implementations will cause significant (over 30 percent) market share loss to IBM and NCR's Teradata (0.3 probability).

Oracle

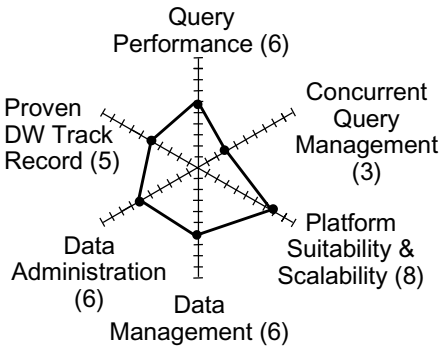
Strengths

- Market leadership and mind share
- Performs well on SMP and NUMA
- Supports tool and BI vendors
- Implements midsize data marts and data warehouses with low concurrency

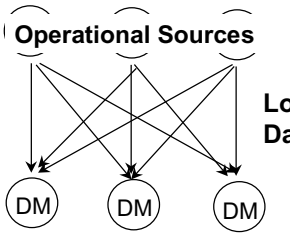
Challenges

- Lack of effectiveness on DMPP platforms
- Concurrent query performance
- Query optimization for complex data models
- Data partitioning
- Manageability
- Unproven with increasing size and complexity

Oracle for Sun



Common Topology



Lots of Data Marts

Gartner

Source: Gartner Research

Conclusion: Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure.

Oracle's greatest strength is its excellent mind share among prospects and with a wide choice of Unix hardware partners. Oracle 8i introduced a number of improvements for star schema data marts and data management (e.g., new partitioning capabilities) that demonstrate slow but continuing DSS maturation. However, at the high end, Oracle is still at a disadvantage to DB2 UDB and Teradata. Oracle will continue to win a major share of the DW DBMS business, since it now has a very capable product for very-large (one terabyte-plus) star schema data marts and large DWs. Nevertheless, Oracle 8i still has high-end data volume and data model complexity (query optimization) challenges, and complex terabyte DWs are not yet commonplace. Most organizations use denormalization techniques to assist the Oracle RDBMS in performing query optimization and reduce workload management challenges. We expect that Oracle 9i will be released in midyear 2001 (0.7 probability) and provide some evolution to Oracle's capabilities, but it will not be revolutionary and will still measurably lag the capabilities of DB2 UDB and Teradata. In the area of concurrent query and mixed-workload management, Oracle 9i provides some focus on these challenges, but the degree to which this helps will be difficult to measure until implementation on production DWs is performed.

Strategic Planning Assumptions: Through 2005, NCR's Teradata will maintain its high-end data warehouse capability lead, which will be much more difficult to defend and lacks the marketing prowess to change the market rules and promote Teradata's midrange capabilities (0.8 probability). By 2005, NCR's Teradata will fall behind the other major DBMS products (DB2 UDB and Oracle) in large and very-large complex data-warehouse capabilities (0.2 probability).

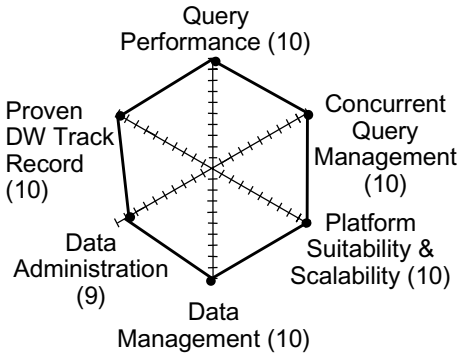
NCR's Teradata

Strengths

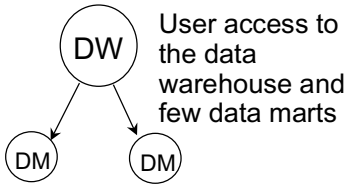
- **Performance standard for very-large data warehouses and data marts**
- **Manageability — low number of DBAs required**
- **Query optimization support for complex data models**
- **Support for concurrent query workloads**

Challenges

- **Lack of platform choice (confusion)**
- **Delays on NT/MPP**
- **Ability to execute well**
- **Ability to keep current capability lead**
- **Marketing and positioning for broader market**
- **Higher initial cost of solution**
- **BI tool and application support**



Common Topology



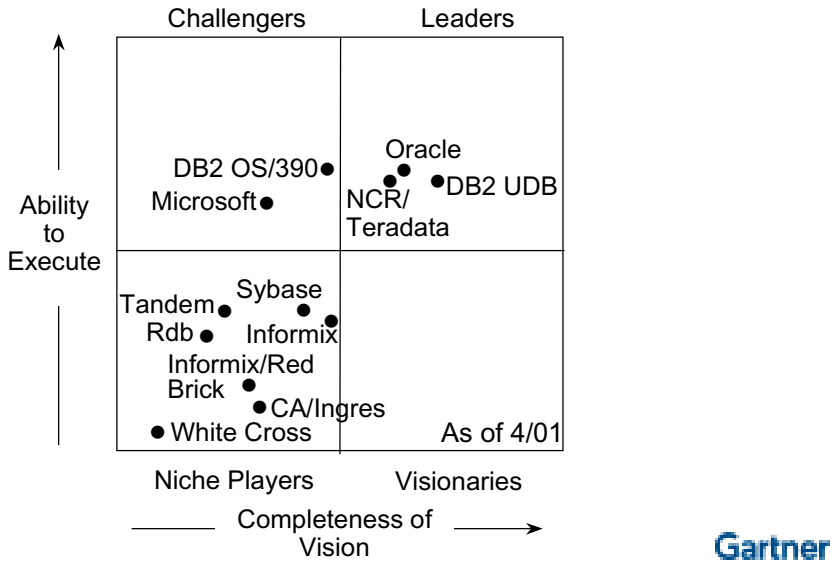
Gartner

Source: Gartner Research

Conclusion: Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure. Clearly, the strength of Teradata has been its proven capability at the high-end and complex (data model and query) DW implementations. Approximately three-and-a-half years ago, there was significant confusion in NCR's DW strategy, when NCR was willing to sell a partner product (e.g., Oracle or Informix) instead of Teradata if that was what the user requested. The solution came in the form of instructing the sales force to sell Teradata at all costs and increasing the investment in the Teradata technology. With the recent creation of the Teradata division, a major divisional alignment within NCR, the enterprise is re-emphasizing the importance of its DW strategy and the Teradata RDBMS. This apparently charges the "autonomous" division with aggressively marketing the Teradata RDBMS on non-NCR hardware platforms, both at the midmarket range (NT SMP) and at the high-end (both NT MPP and Solaris). The introduction of Teradata on non-NCR hardware puts NCR in the middle of a paradox that will be difficult to address successfully. To date, NCR's Teradata division has only made an extremely limited marketing effort in its support for non-NCR hardware, and delivered product with support for NT SMP only. Solaris and NT MPP, supported with non-NCR hardware, remain elusive. Although NCR is not a Unix server leader, sales of its hardware (due to Teradata) have justified its continued presence in the server market. Active warehouse and customer-relationship-management applications have become a focus.

Strategic Planning Assumptions: Through 2005, unlike the case of OLTP scalability, there will be significant differences between the capabilities of the various DBMS products and their ability to effectively support large data-warehouse implementations (0.7 probability). By 2003, the leading RDBMS products will be virtually equal in their capability to support large and very-large data-warehouse implementations (0.3 probability).

Data Warehouse DBMS Magic Quadrant



Source: Gartner Research

Conclusion: Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure. The DBMS market supporting data warehousing and complex DSS applications continues to consolidate around IBM, NCR and Oracle, with Microsoft at the low end of the fray. Most enterprises continue to see Unix as the natural home for a scalable DW, unless there is strong mainframe culture — and in those few situations — DB2 for OS/390 is preferred as the natural platform. NCR's Teradata continues to make slow but steady progress in gaining market share. The ongoing confusion and lack of clarity related to its portability strategy (or lack thereof) has caused some organizations to shy away from the product. IBM is having success with DB2 UDB on the IBM RS/6000, and it is gaining perception as a viable product for complex DW implementations. Leveraging its dominance of the Unix RDBMS market, Oracle continues to garner large-DW-related revenue, despite doubts over its technical capability and its capability to tackle the high end. On NT, the focus has remained on small-to-midsize data marts with little complexity; but little user attention has been given to the scalability of NT clusters, and few midsize-to-large DW implementations exist. Microsoft will have a growing impact, and SQL Server 2000 provides incremental capabilities, but NT platform and RDBMS functionality and maturity issues will limit its range. Both Informix and Sybase garner little attention in the data warehousing market space, and their respective DBMS products have fallen back technologically.

Recommendations

- **Leverage existing skill sets of application support teams when preparing data for a data warehouse.**
- **Where possible, select tools for data-warehouse data preparation that can leverage other data integration efforts that the enterprise is undertaking.**
- **Selecting ETT tools needs to be treated as a tactical selection with an ROI documented, measured and verified.**
- **Do not assume that the success of your OLTP DBMS will guarantee success in your data warehouse implementation. Carefully evaluate its capabilities against other leading DBMS products.**

Source: Gartner Research

Conclusion: Not all RDBMS technology is created equal, and a thorough evaluation of requirements is necessary to reduce the risk of failure.

Here, we present the next steps to be followed in simplifying the complexity of preparing data for the DW, depending on where the enterprise is today with regard to deploying a DW infrastructure. For those that do not yet have a DW architecture, the first decision will be whether to buy a "best of market" offering or custom build the DW architecture using a set of "best of breed" point products. Best-of-market offerings are now available from ERP vendors and third parties (e.g., Acta Technology; DecisionPoint Applications; Shell Services International; Fair, Isaac; Sybase and many other industry or functionally oriented packaged environments). These should be evaluated on the strength, scalability and extensibility of their infrastructures, and on their BI functionality. Although all of the major RDBMS vendors offer products for building a DW architecture, most of their offerings are still individual point products rather than integrated technology suites. Organizations pursuing best-of-breed, custom-built approaches should consider some of the ETT and data quality tools reviewed here. Organizations that already have a DW in place, but are still preparing the data by hand, should evaluate the ETT functionality that is now available from their DBMS vendor or BI tool vendor in comparison to leading third-party point products.

This is the final
page of the presentation

Click to continue...