A 101communications Publication

TDWI

REPORT SERIES

Evaluating ETL and Data Integration Platforms

Wayne Eckerson and Colin White

tdwi
THE DATA WAREHOUSING INSTITUTE

Research Sponsors

Business Objects

DataMirror Corporation

Hummingbird Ltd

Informatica Corporation

Acknowledgements

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who responded to our requests for phone interviews. Second, we'd like to thank Steve Tracy and Pieter Mimno who reviewed draft manuscripts and provided feedback, as well as our report sponsors who reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: Denelle Hanlon, Theresa Johnston, Marie McFarland, and Donna Padian.

About TDWI

The Data Warehousing Institute (TDWI), a division of 101communications LLC, is the premier provider of in-depth, high-quality education and training in the business intelligence and data warehousing industry. TDWI supports a worldwide membership program, quarterly educational conferences, regional seminars, onsite courses, leadership awards programs, numerous print and online publications, and a public and private (Members-only) Web site.

TDWI
REPORT SERIES

Evaluating ETL and Data Integration Platforms

by Wayne W. Eckerson and Colin White

Table of Contents

About the Authors

WAYNE ECKERSON is director of research for The Data Warehousing Institute (TDWI), the leading provider of high-quality, in-depth education and research services to data warehousing and business intelligence professionals worldwide. Eckerson oversees TDWI's Member publications and research services.

Eckerson has written and spoken on data warehousing and business intelligence since 1994. He has published in-depth reports and articles about data quality, data warehousing, customer relationship management, online analytical processing (OLAP), Web-based analytical tools, analytic applications, and portals, among other topics. In addition, Eckerson has delivered presentations at industry conferences, user group meetings, and vendor seminars. He has also consulted with many vendor and user firms.

Prior to joining TDWI, Eckerson was a senior consultant at the Patricia Seybold Group, and director of the Group's Business Intelligence & Data Warehouse Service, which he launched in 1996.

COLIN WHITE is president of Intelligent Business Strategies. White is well known for his in-depth knowledge of leading-edge business intelligence, enterprise portal, database and Web technologies, and how they can be integrated into an IT infrastructure for building and supporting the intelligent business.

With more than 32 years of IT experience, White has consulted for dozens of companies throughout the world and is a frequent speaker at leading IT events. He is a faculty member at TDWI and currently serves on the advisory board of TDWI's Business Intelligence Strategies program. He also chairs a portals and Web Services conference.

White has co-authored several books, and has written numerous articles on business intelligence, portals, database, and Web technology for leading IT trade journals. He writes a regular column for *DM Review* magazine entitled "Intelligent Business Strategies." Prior to becoming an analyst and consultant, White worked at IBM and Amdahl.

About the TDWI Report Series

The TDWI Report Series is designed to educate technical and business professionals about critical issues in BI. TDWI's in-depth reports offer objective, vendor-neutral research consisting of interviews with industry experts and a survey of BI professionals worldwide. TDWI in-depth reports are sponsored by vendors who collectively wish to evangelize a discipline within BI or an emerging approach or technology.

TDWI Definitions

ETL - Extract, transform, and load (ETL) tools play a critical part in creating data warehouses, which form the bedrock of business intelligence (see below). ETL tools sit at the intersection of myriad source and target systems and act as a funnel to pull together and blend heterogeneous data into a consistent format and meaning and populate data warehouses.

Business Intelligence - TDWI uses the term "business intelligence" or BI as an umbrella term that encompasses ETL, data warehouses, reporting and analysis tools, and analytic applications. BI projects turn data into information, knowledge, and plans that drive profitable business decisions.

Scope and Methodology

Report Scope. This report examines the current and future state of ETL tools. It describes how business requirements are fueling the creation of a new generation of products called data integration platforms. It examines the pros and cons of building versus buying ETL functionality and assesses the challenges and success factors involved with implementing ETL tools. It finishes by providing a list of evaluation criteria to assist organizations in selecting ETL products or validating current ones.

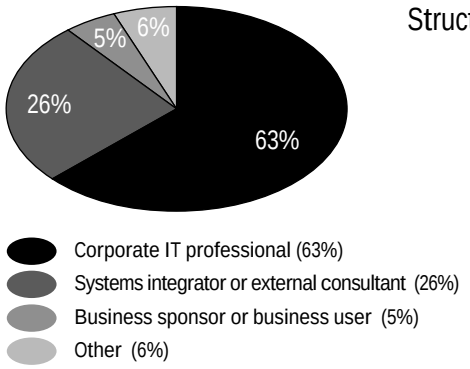
Methodology. The research for this report was conducted by interviewing industry experts, including consultants, industry analysts, and IT professionals, who have implemented ETL tools. The research is also based on a survey of 1000+ business intelligence professionals that TDWI conducted in November 2002.

Survey Methodology. TDWI received 1,051 responses to the survey. Of these, TDWI qualified 741 respondents who had both deployed ETL functionality and were either IT professionals or consultants at end-user organizations. Branching logic accounts for the variation in the number of respondents to each question. For example, respondents who "built" ETL programs were asked a few different questions from those who "bought" vendor ETL tools. Multi-choice questions and rounding techniques account for totals that don't equal 100 percent.

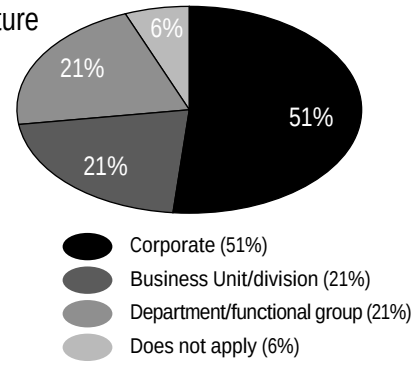
Survey Demographics. Most respondents were corporate IT professionals who work at large U.S. companies in a range of industries. (See the illustrations on this page for breakouts.)

Demographics

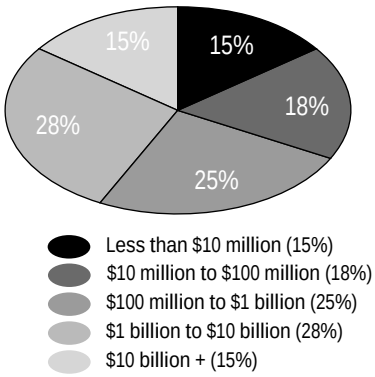
Position



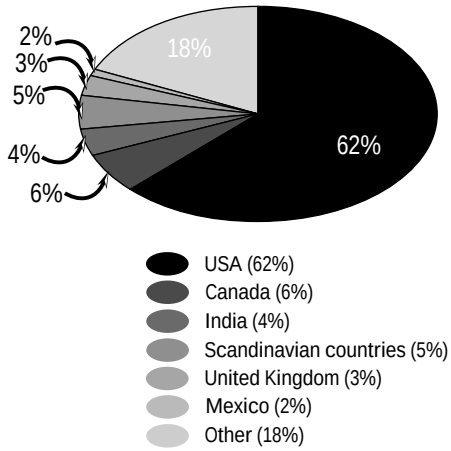
Organization Structure



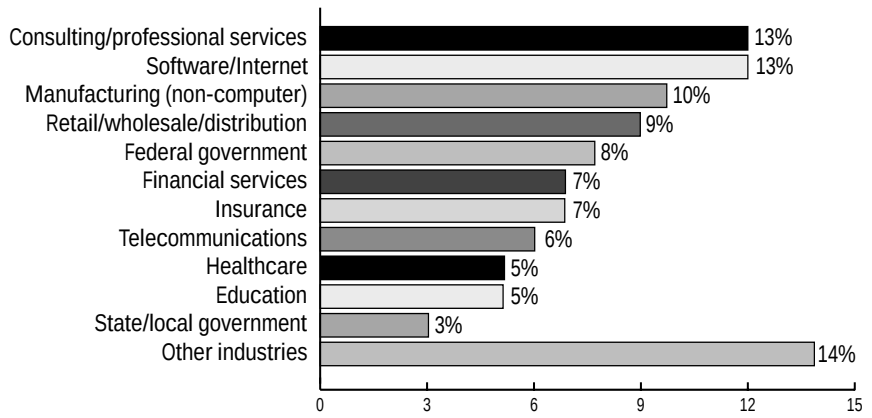
Company Revenues



Country



Industry



Executive Summary: The Role of ETL in Business Intelligence

ETL: The Heart and Soul of BI

ETL is the heart and soul of business intelligence (BI). ETL processes bring together and combine data from multiple source systems into a data warehouse, enabling all users to work off a single, integrated set of data—a single version of the truth. The result is an organization that no longer spins its wheels collecting data or arguing about whose data is correct, but one that uses information as a key process enabler and competitive weapon.

In these organizations, BI systems are no longer *nice to have*, but essential to success. These systems are no longer stand-alone and separate from operational processing—they are integrated with overall business processes. As a result, an effective BI environment based on integrated data enables users to make strategic, tactical, and operational decisions that drive the business on a daily basis.

ETL Work Consumes 60 to 80 Percent of all BI Projects

Why ETL is Hard. According to most practitioners, ETL design and development work consumes 60 to 80 percent of an entire BI project. With such an inordinate amount of resources tied up in ETL work, it behooves BI teams to optimize this layer of their BI environment.

ETL is so time consuming because it involves the unenviable task of re-integrating the enterprise's data from scratch. Over the span of many years, organizations have allowed their business processes to *dis-integrate* into dozens or hundreds of local processes, each managed by a single fiefdom (e.g., departments, business units, divisions) with its own systems, data, and view of the world.

With the goal of achieving a single version of truth, business executives are appointing BI teams to *re-integrate* what has taken years or decades to undo. Equipped with ETL and modeling tools, BI teams are now expected to swoop in like conquering heroes and rescue the organization from information chaos. Obviously, the challenges and risks are daunting.

ETL Tools Are the Most Important in a BI Toolbox

Mitigating Risk. ETL tools are perhaps the most critical instruments in a BI team's toolbox. Whether built or bought, a good ETL tool in the hands of an experienced ETL developer can speed deployment, minimize the impact of systems changes and new user requirements, and mitigate project risk. A weak ETL tool in the hands of an untrained developer can wreak havoc on BI project schedules and budgets.

ETL in Flux

Given the demands placed on ETL and the more prominent role that BI is playing in corporations, it is no wonder that this technology is now in a state of flux.

Users Seek Integrated Toolsets

More Complete Solutions. Organizations are now pushing ETL vendors to deliver more complete BI "solutions." Primarily, this means handling additional back-end data management and processing responsibilities, such as providing data profiling, data cleansing, and enterprise meta data management utilities. A growing number of users also want BI vendors to deliver a complete solution that spans both back-end data management functions and front-end reporting and analysis applications.

ETL Tools Must Process More Data in Less Time

Better Throughput and Scalability. They also want ETL tools to increase throughput and performance to handle exploding volumes of data and shrinking batch windows. Rather than refresh the entire data warehouse from scratch, they want ETL tools to capture and update changes that have occurred in source systems since the last load.

More Sources, Greater Complexity, Better Administration. ETL tools also need to handle a wider variety of source system data, including Web, XML, and packaged applications. To integrate these diverse data sets, ETL tools must also handle more complex mappings and transformations, and offer enhanced administration to improve reliability and speed deployments.

Users Want Timelier Data

"Near-Real-Time" Data. Finally, ETL tools need to feed data warehouses more quickly with more up-to-date information. This is because batch processing windows are shrinking and business users want integrated data delivered on a timelier basis (i.e., the previous day, hour, or minute) so they can make critical operational decisions without delay.

Clearly, the market for ETL tools is changing and expanding. In response to user requirements, ETL vendors are transforming their products from single-purpose ETL products into multi-purpose *data integration platforms*. BI professionals need help understanding these market and technology changes as well as how to leverage the convergence of ETL with new technologies to optimize their BI architectures and ensure a healthy return on their ETL investments.

The Evolution of ETL

Framework and Components

Before we examine the future of ETL, we will define the major components in an ETL framework.

ETL stands for *extract, transform, and load*. That is, ETL programs periodically extract data from source systems, *transform* the data into a common format, and then load the data into the target data store, usually a data warehouse.

Do We Need a New Acronym?

As an acronym, however, ETL only tells part of the story. ETL tools also commonly *move* or transport data between sources and targets, *document* how data elements change as they move between source and target (i.e., meta data), *exchange* this meta data with other applications as needed, and *administer* all run-time processes and operations (e.g., scheduling, error management, audit logs, and statistics). A more accurate acronym might be EMTDLEA!

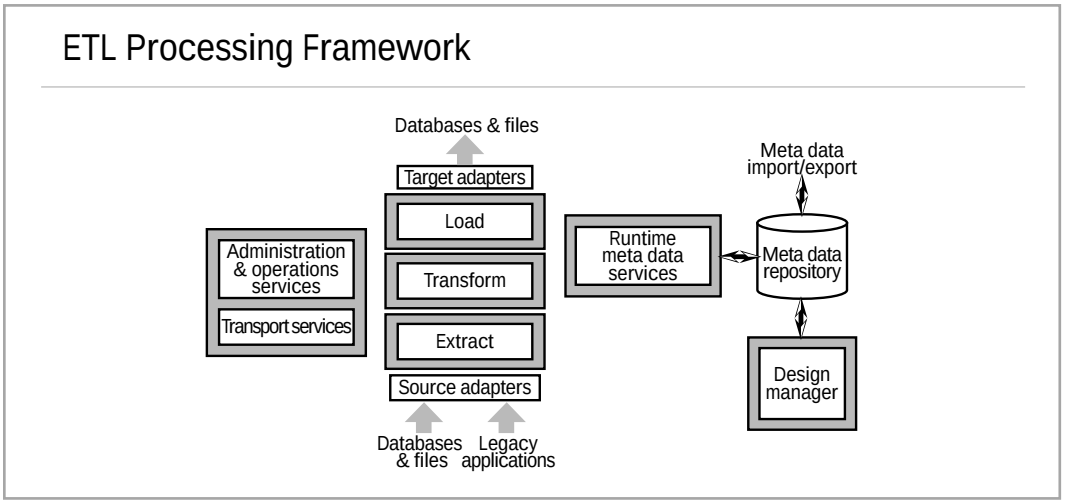


Illustration S1. The diagram shows the core components of an ETL product. Courtesy of Intelligent Business Strategies.

The diagram on page 5 depicts the major components involved in ETL processing. The following bullets describe each component in more detail:

- **Design manager:** Provides a graphical mapping environment that lets developers define source-to-target mappings, transformations, process flows, and jobs. The designs are stored in a meta data repository.
- **Meta data management:** Provides a repository to define, document, and manage information (i.e., meta data) about the ETL design and runtime processes. The repository makes meta data available to the ETL engine at run time and other applications
- **Extract:** Extracts source data using adapters, such as ODBC, native SQL formats, or flat file extractors. These adapters consult meta data to determine which data to extract and how.
- **Transform:** ETL tools provide a library of transformation objects that let developers transform source data into target data structures and create summary tables to improve performance.
- **Load:** ETL tools use target data adapters, such as SQL or native bulk loaders, to insert or modify data in target databases or files.
- **Transport services:** ETL tools use network and file protocols (e.g., FTP) to move data between source and target systems and in-memory protocols (e.g., data caches) to move data between ETL run-time components.
- **Administration and operation:** ETL utilities let administrators schedule, run, and monitor ETL jobs as well as log all events, manage errors, recover from failures, and reconcile outputs with source systems.

The components above come with most vendor-supplied ETL tools, but they can also be built by data warehouse developers. A majority of companies have a mix of packaged and homegrown ETL applications. In some cases, they use different tools in different projects, or they use custom code to augment the functionality of an ETL product. The pros and cons of building and buying ETL components are discussed later in the report. (See “Build or Buy?” on page 15.)

A Quick History

Code Generation Tools

In the early 1990s, most organizations developed custom code to extract and transform data from operational systems and load it into data warehouses. In the mid-1990s, vendors recognized an opportunity and began shipping ETL tools designed to reduce or eliminate the labor-intensive process of writing custom ETL programs.

Early vendor ETL tools provided a graphical design environment that generated third-generation language (3GL) programs, such as a COBOL. Although these early code-generation tools simplified ETL development work, they did little to automate the runtime environment or lessen code maintenance and change control work. Often, administrators had to manually distribute and manage compiled code, schedule and run jobs, or copy and transport files.

ETL Engines Interpret Design Rules at Runtime

Engine-Based Tools

To automate more of the ETL process, vendors began delivering “engine-based” products in the mid to late 1990s that employed proprietary scripting languages running within an ETL or DBMS server. These ETL engines use language interpreters to process ETL workflows at runtime. The ETL workflows defined by developers in the graphical environment are stored in a meta data repository, which the engine reads at runtime to determine how to process incoming data.

Although this interpretive approach more tightly unifies design and execution environments, it doesn’t necessarily eliminate all custom coding and maintenance. To handle complex or unique requirements, developers often resort to coding custom routines and exits that the tool

accesses at runtime. These user-developed routines and exits increase complexity and maintenance, and therefore should be used sparingly.

ETL Engines Centralize Processing in a High-Performance Server

All Processing Occurs in the Engine. Another significant characteristic of an engine-based approach is that all processing takes place in the engine, not on source systems (although hybrid architectures exist). The engine typically runs on a Windows or UNIX machine and establishes direct connections to source systems. If the source system is non-relational, administrators use a third-party gateway to establish a direct connection or create a flat file to feed the ETL engine.

Some engines support parallel processing, which enables them to process ETL workflows in parallel across multiple hardware processors. Others require administrators to manually define a parallel processing framework in advance (using partitions, for example)—a much less dynamic environment.

Code Generators versus Engines

There is still a debate about whether ETL engines or code generators, which have improved since their debut in the mid-1990s, offer the best functionality and performance.

Benefits of Code Generators. "The benefit of code generators is that they can handle more complex processing than their engine-based counterparts," says Steve Tracy, assistant director of information delivery and application strategy at Hartford Life Insurance in Simsbury, CT. "Many engine-based tools typically want to read a record, then write a record, which limits their ability to efficiently handle certain types of transformations."

Code Generators Eliminate the Need to Maintain Custom Routines and Exits

Consequently, code generators also eliminate the need for developers to maintain user-written routines and exits to handle complex transformations in ETL workflows. This avoids creating "blind spots" in the tools and their meta data.

In addition, code generators produce compiled code to run on various platforms. Compiled code is not only fast, it also enables organizations to distribute processing across multiple platforms to optimize performance.

"Because the [code generation] product ran on multiple platforms, we did not have to buy a huge server," says Kevin Light, managing consultant of business intelligence solutions at EDS Canada. "Consequently, our client's overall capital outlay [for a code generation product] was less than half what it would cost them to deploy a comparable engine-based product."

Benefits of Engines. Engines, on the other hand, concentrate all processing in a single server. Although this can become a bottleneck, administrators can optimally configure the engine platform to deliver high performance. Also, since all processing occurs on one machine, administrators can more easily monitor performance and quickly upgrade capacity if needed to meet system level agreements.

This approach also removes political obstacles that arise when ETL administrators try to distribute transformation code to run on various source systems. By offloading transformation processing from source systems, ETL engines interfere less with critical runtime operations.

Visual Development Environments Help Manage Complex Jobs

Enhanced Graphical Development. In addition, most engine-based ETL products offer visual development environments that make them easier to use. These graphical workspaces enable developers to create ETL workflows comprised of multiple ETL objects for defining data mappings and transformations. (See Illustration S2.) Although code generators also offer visual development environments, they often aren't as easy to use as engine-based tools, lengthening overall development times, according to many practitioners.

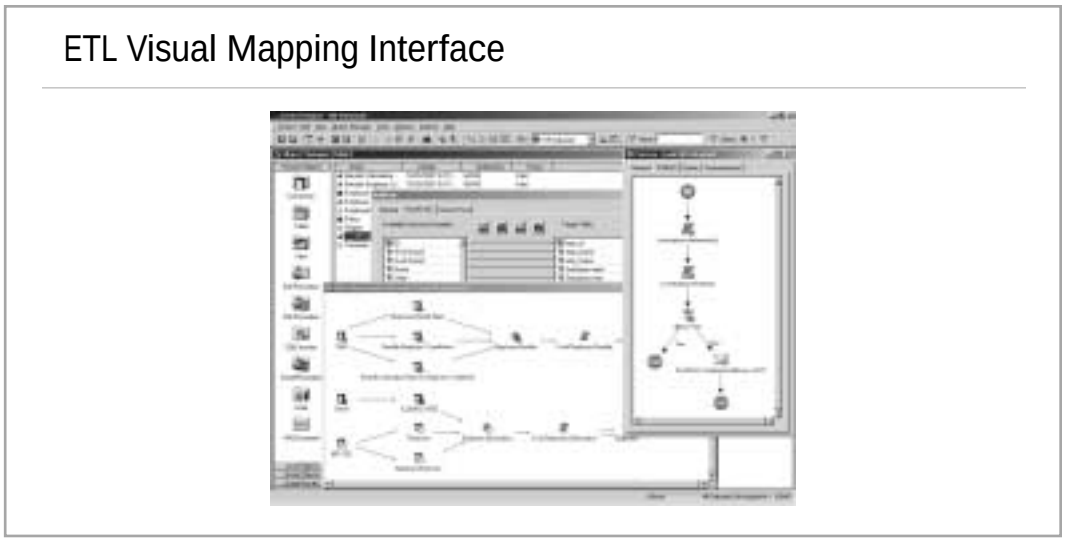


Illustration S2. Many ETL tools provide a graphical interface for mapping sources to targets and managing complex workflows.

Database-Centric ETL

Today, several DBMS vendors embed ETL capabilities in their DBMS products (as well as OLAP and data mining capabilities). Since these vendors offer ETL capabilities at little or no extra charge, organizations are seriously exploring this option because it promises to reduce costs and simplify their BI environments.

"Why Purchase an ETL Tool When Our Database Bundles One?"

In short, users are now asking, "Why should we purchase a third-party ETL tool when we can get ETL capabilities for free from our database vendor of choice? What are the additional benefits of buying a third-party ETL tool?"

To address these questions, organizations first need to understand the level and type of ETL processing that database vendors support. Today, there are three basic groups:

Cooperative ETL. Here, third-party ETL tools can leverage common database functionality to perform certain types of ETL processing. For example, third-party ETL tools can leverage stored procedures and enhanced SQL to perform transformations and aggregations in the database where appropriate. This enables third-party ETL tools to optimize performance by exploiting the optimization, parallel processing, and scalability features of the DBMS. It also improves recoverability since stored procedures are maintained in a common recoverable data store.

Complementary ETL. Some database vendors now offer ETL functions that mirror features offered by independent ETL vendors. For example, materialized views create summary tables that can be automatically updated when detailed data in one or more associated tables change. Summary tables can speed up query performance by several orders of magnitude. In addition, some DBMS vendors can issue SQL statements that interact with Web Services-based applications or messaging queues, which are useful when building and maintaining near-real-time data warehouses.

Competitive ETL. Most database vendors offer graphical development tools that exploit the ETL capabilities of their database products. These tools provide many of the features offered by third-party ETL tools, but at zero cost, or for a license fee that is a fraction of the price of independent tools. At present, however, database-centric ETL solutions vary considerably in quality and functionality. Third-party ETL tools still offer important advantages, such as the

range of data sources supported, transformation power, and administration. Today, database-centric ETL products are useful for building departmental data marts, but this will change over time as DBMS vendors enhance their ETL capabilities.

Database Vendors Both Enhance and Compete with ETL Vendors

In summary, database vendors currently offer ETL capabilities that both enhance and compete with independent ETL tools. We expect database vendors to continue to enhance their ETL capabilities and compete aggressively to increase their share of the ETL market.

Data Integration Platforms

During the past several years, business requirements for BI projects have expanded dramatically, placing new demands on ETL tools. These requirements are outlined in the following section (See "*ETL Trends and Requirements*"). Consequently, ETL vendors have begun to dramatically change their products to meet these requirements. The result is a new generation ETL tool that TDWI calls a ***data integration platform***.

Higher Performance, Complete Solution

As a **platform**, this emerging generation of ETL tools provides greater performance, throughput, and scalability to process larger volumes of data at higher speeds. To deal with shrinking batch windows, these platforms also load data warehouses more quickly and reliably, using change data capture techniques, continuous processing, and improved runtime operations. The platforms also provide expanded functionality, especially in the areas of data quality, transformation power, and administration.

As a **data integration** hub, these products connect to a broader array of databases, systems, and applications as well as other integration hubs. They capture and process data in batch or real time using either a hub-and-spoke or peer-to-peer information delivery architecture. They coordinate and exchange meta data among heterogeneous systems to deliver a highly integrated environment that is easy to use and adapts well to change.

ETL Vendors Are Expanding Vertically and Horizontally

Solution Sets. Ultimately, data integration platforms are designed to meet a larger share of an organization's BI requirements. To do this, some ETL vendors are extending their product lines *horizontally*, adding data quality tools and near-real-time data capture facilities to provide a complete data management solution. Others are extending *vertically*, adding analytical tools and applications to provide a complete BI solution.

ETL versus EAI

To be clear, ETL tools are just one of several technologies vying to become the preeminent enterprise data integration platform. One of the most important technologies is enterprise application integration (EAI) software, such as BEA Systems, Inc.'s WebLogic Platform and Tibco Software Inc.'s ActiveEnterprise.

EAI software enables developers to create real-time, event-driven interfaces among disparate transaction systems. For example, during the Internet boom, companies flocked to EAI tools to connect e-commerce systems with back-end inventory and shipping systems to ensure that Web sites accurately reflected product availability and delivery times.

Some ETL and EAI Vendors Are Partnering...

Although EAI software is event-driven and supports transaction-style processing, it does not generally have the same transformation power as an ETL tool. On the other hand, most ETL tools do not support real-time data processing. To overcome these limitations, some ETL and EAI vendors are now partnering to provide the best of both approaches. EAI software captures data and application events in real time and passes them to the ETL tools, which transform the data and loads it into the BI environment.

ETL and EAI Will Converge

Already, some ETL vendors have begun incorporating EAI functionality into their core engines. These tools contain adapters to operational applications and operate in an "always awake" mode so they can receive and process data and events generated by the applications in real time.

Future Reality. It won't be long before all ETL and EAI vendors follow suit, integrating both ETL and EAI capabilities within a single product. The winners in this race will have a substantial advantage in vying for delivery of a complete, robust data integration platform.

ETL Trends and Requirements

As mentioned earlier, business requirements are fueling the evolution of ETL into data integration platforms. Business users are demanding access to more timely, detailed, and relevant information to inform critical business decisions and drive fast-moving business processes.

Large Volumes of Data

More History, More Detail, More Subject Areas

More business users today want access to more granular data (e.g., transactions) and more years of history across more subject areas than ever before. Not surprisingly, data warehouses are exploding in size and scope. Terabyte data warehouses—a rarity several years ago—are now fairly commonplace. The largest data warehouses exceed 50 terabytes of raw data, and the size of these data warehouses is putting enormous pressure on ETL administrators to speed up ETL processing.

Our survey indicates that the percentage of organizations loading more than 500 gigabytes will triple in 18 months, while those loading less than 1 gigabyte will decrease by a third. (See Illustration 1.)

Eighteen-Month Plans for Data Load Volumes

	TODAY	18 MONTHS
Less than 1GB	59%	40%
1GB to 500GB	38%	50%
500GB+	3%	10%

Illustration 1. The average data load will increase significantly in the next 18 months. Based on 756 respondents.

Diverse Data Sources

On Average, Data Warehouses Cull Data from 12 Distinct Sources

One reason for increased data volumes is that business users want the data warehouse to cull data from a wider variety of systems. According to our survey, on average, organizations now extract data from 12 distinct data sources. This average will inexorably increase over time as organizations expand their data warehouses to support more subject areas and groups in the organization.

Although almost all companies use ETL to extract data from relational databases, flat files, and legacy systems, a significant percentage now want to extract data from application packages, such as SAP R/3 (39 percent), XML files (15 percent), Web-based data sources (15 percent), and EAI software (12 percent). (See Illustration 2.)

Excel Spreadsheets Are a Critical Data Source

"Spreadmarts." In addition, many respondents also wrote in the "other" category that they want ETL tools to extract data from Microsoft Excel and Access files. In many companies, critical data

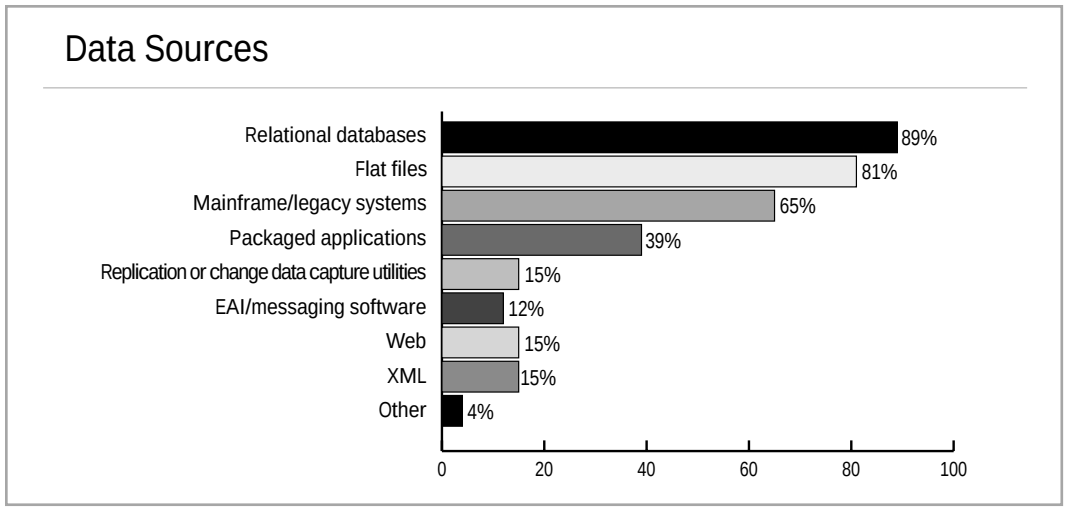


Illustration 2. Types of data sources that ETL programs process. Multi-choice question, based on 755 respondents.

is locked up in these personal data stores, which are controlled by executives or managers running different divisions or departments. Each group populates these data stores from different sources at different times using different rules and definitions. As a result, these spreadsheets become surrogate independent data marts—or "spreadmarts" as TDWI calls them. The proliferation of spreadmarts precipitates organizational feuds ("dueling spreadsheets") that can paralyze an organization and drive a CEO to the brink of insanity.

Shrinking Batch Windows

More Data, Less Time

With the expansion in data volumes, many organizations are finding it virtually impossible to load a data warehouse in a single batch window. There are no longer enough hours at night or on the weekend to finish loading data warehouses containing hundreds of gigabytes or terabytes of data.

24x7 Requirements. Compounding the problem, batch windows are shrinking or becoming non-existent. This is especially true in large organizations whose data warehouses and operational systems span regions or continents and must be available around the clock. There is no longer any "systems downtime" to execute time consuming extracts and loads.

Data Has Different "Expiration Dates"

Variable Update Cycles. Finally, the growing number and diversity of sources that feed the data warehouse make it difficult to coordinate a single batch load. Each source system operates on a different schedule and each contains data with different levels of "freshness" or relevancy for different groups of users. For example, retail purchasing managers find little value in sales data if they can't view it within one day, if not immediately. Accountants, however, may want to examine general ledger data at the end of the month.

Operational Decision Making

Users Want Timelier Data

Another reason that ETL tools must support continuous extract/load processing is that users want timelier data. Data warehouses traditionally support strategic or tactical decision making based on historical data compiled over several months or years. Typical strategic or tactical questions are:

- *"What are our year-over-year trends?"*
- *"How close are we to meeting this month's revenue goals?"*
- *"If we reduce prices, what impact will it have on our sales and inventory for the next three months?"*

Operational BI. However, business users increasingly want to make operational decisions based on yesterday's or today's data. They are now asking:

- *"Based on yesterday's sales, which products should we ship to which stores at what prices to maximize today's revenues?"*
- *"What is the most profitable way to reroute packages off a truck that broke down this morning?"*
- *"What products and discounts should I offer to the customer that I'm currently speaking with on the phone?"*

Near-real-time Loads Will Triple in 18 Months

According to our survey, more than two-thirds of organizations already load their data warehouses on a nightly basis. (See Illustration 3.) However, in the next 18 months, the number of organizations that will load their data warehouses multiple times a day will double, and the number that will load data warehouses in near real time will triple!

Data Warehouse Load Frequency

	TODAY	IN 18 MONTHS
Monthly	32%	27%
Weekly	34%	29%
Daily/nightly	69%	65%
Multiple times per day	15%	30%
Near real time	6%	19%

Illustration 3. Based on 754 respondents.

To meet operational decision-making requirements, ETL processing will need to support two types of processing: (1) large batch ETL jobs that involve extracting, transforming, and loading a significant amount of data, and (2) continuous processing that captures data and application events from source systems and loads them into data warehouses in rapid intervals or near real time.

Most organizations will need to merge these two types of processing, which is complicated because of the numerous interdependencies that administrators must manage to ensure consistency of information in the data warehouse and reports.

Data Quality Add-Ons

Many organizations want to purchase complete solutions, not technology or tools for building infrastructure and applications.

Users Want ETL Vendors to Offer Data Quality Tools

Add-on Products. When asked which add-on products from ETL vendors are "very important," users expressed significant interest in data cleansing and profiling tools. (See Illustration 4.) This is not surprising since many BI projects fail because the IT department or (more than likely) an outside consultancy underestimates the quality of source data.

Code, Load, and Explode. A common scenario is the "code, load, and explode" phenomenon. This is where ETL developers code the extracts and transformations, then start processing only to discover an unacceptably large number of errors due to unanticipated values in the source data files. They fix the errors, rerun the ETL process, only to find more errors, and so on. This ugly scenario repeats itself until the project deadlines and budgets become imperiled and angry business sponsors halt the project.

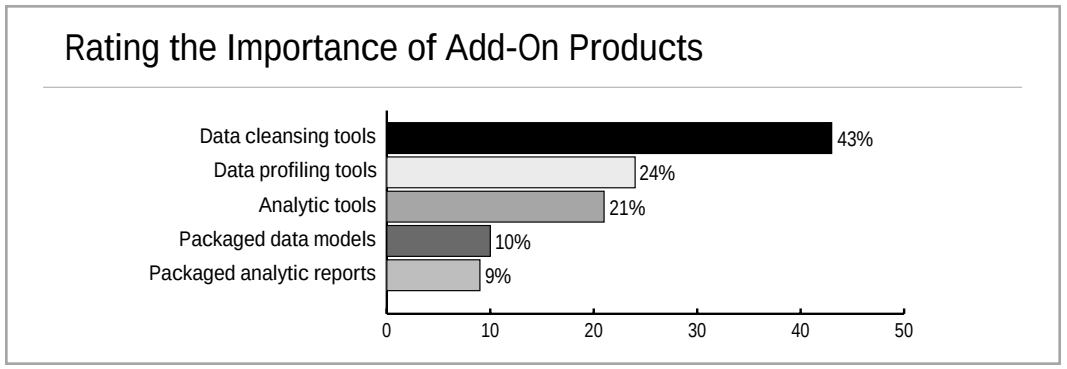


Illustration 4. A significant percentage of users want ETL tools to offer data cleansing and profiling capabilities. Based on 740 respondents.

Lack of Knowledge about Source Data Cripples Projects

"In my experience, 10 percent [of the time and cost associated with ETL] is learning the ETL tool the first time out and making it conform to your will," said John Murphy, an IT professional in a newsgroup posting. "Ninety percent is the interminable iterative process of learning just how little you know about the source data, how dirty that data really is, how difficult it is to establish a cleansing policy that results in certifiably correct data on the target that matches the source, and how long it takes to re-work all the assumptions you made at the outset while you were still an innocent."

ETL Tools Assume the Data Is Clean

The problem with current ETL tools is that they *assume* the data they receive is clean and consistent. They can't assess the consistency or accuracy of source data and they can't handle specialized cleansing routines, such as name and address scrubbing. To meet these needs, ETL vendors are partnering with or acquiring data quality vendors to integrate specialized data cleansing routines within ETL workflows.

Meta Data Management

As the number and variety of source and target systems proliferate, BI managers are seeking automated methods to document and manage the interdependencies among data elements in the disparate systems. In short, they want a global meta data management solution.

Users Want ETL Tools to Play a Pivotal Role in Meta Data

As the hub of a BI project, an ETL tool is well positioned to document, manage, and coordinate information about data within data modeling tools, source systems, data warehouses, data marts, analytical tools, portals, and even other ETL tools and repositories. Many users would like to see ETL tools play a more pivotal role in managing global meta data, although others think ETL tools will always be inherently unfit to play the role of global traffic cop.

"Certainly, ETL vendors should store meta data for its own domain, but if they start trying to become a global meta data repository for all BI tools, they will have trouble," says Tracy from Hartford Life. "I'd prefer they focus on improving their performance, manageability, impact analysis reports, and ability to support complex logic."

Enforce Data Consistency. Nevertheless, large companies want a global meta data repository to manage and distribute "standard" data definitions, rules, and other elements within and throughout a network of interconnected BI environments. This global repository could help ensure data consistency and a "single version of the truth" throughout an organization.

Most ETL Tools Fall Short of Expectations

Although ETL vendors often pitch their global meta data management capabilities, most tools today fall short of user expectations. (And even single vendor all-in-one BI suites fall short.) When asked in an open-ended question, "What would make your ETL tool better?" many survey respondents wrote "better meta data management" or "enhanced meta data integration."

"I don't know if ETL tools should be used for a [BI] meta data repository, but no one else is stepping up to the plate," says EDS Canada's Light. Light says the gap between expectation and reality sometimes puts him as a consultant in a tough position. "Many clients think the ETL tools can do it all, and we [as consultants] end up as the meat in a sandwich trying to force a tool to do something it can't."

Packaged Solutions

Besides add-on products, an increasing number of organizations are looking to buy comprehensive, off-the-shelf packages that provide near instant insight to their data. For instance, more than a third of respondents (39 percent) said they are more attracted to vendors that offer ETL as part of a complete application or analytic suite. (See Illustration 5.)

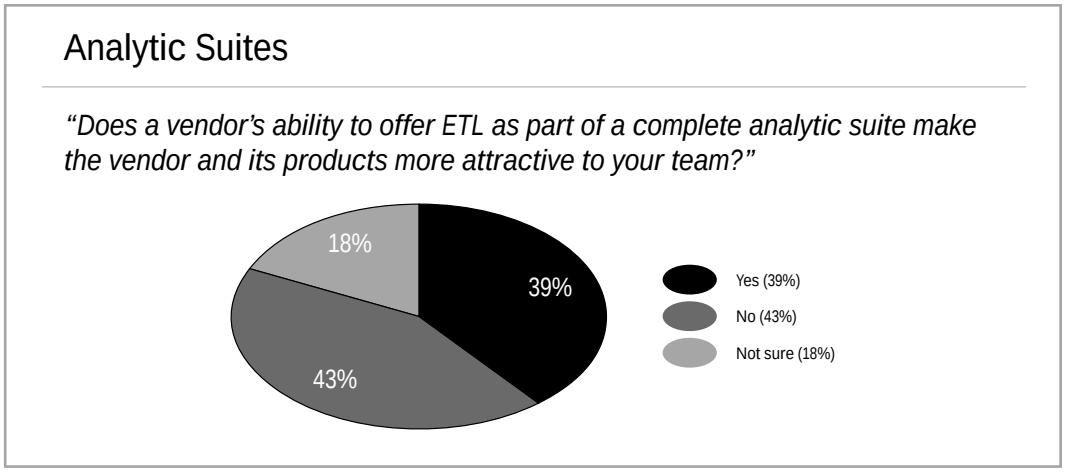


Illustration 5. Analytic suites. Based on 747 respondents.

Packaged Solutions Can Speed Deployment

Kerr-McGee Corporation, an energy and chemicals firm in Oklahoma City, purchased a packaged solution to support its Oracle Financials application. "We had valid numbers within a week, which is incredible," says Brian Morris, a data warehousing specialist at the firm. "We don't have a lot of time so anything that is pre-built is helpful."

A packaged approach excels, as in Kerr-McGee's case, when there is a single source, which is a packaged application itself with a documented data model and API. The package Kerr-McGee used also supplied analytic reports, which the firm did not use since it already had reports defined in another reporting tool.

Some Skepticism. However, there is skepticism among some users about ETL vendors who want to deliver end-to-end BI packages. "We want ETL vendors to focus on their core competency, not their brand," says Cynthia Connolly, vice president of application development at Alliance Capital Management. "We want them to enhance their ETL products, not branch off into new areas."

Enterprise Infrastructure

Organizations Want a Standard Data Integration Infrastructure

Multi-Faceted Usage. Finally, organizations also want to broaden their use of ETL to support additional applications besides BI. (See Illustration 6.) In essence, they want to standardize on a single infrastructure product to support all their enterprise data integration requirements.

A majority of organizations already use ETL tools for BI, data migration, and conversion projects. But, a growing number will use ETL tools to support application integration and master reference data projects in the next 18 months.

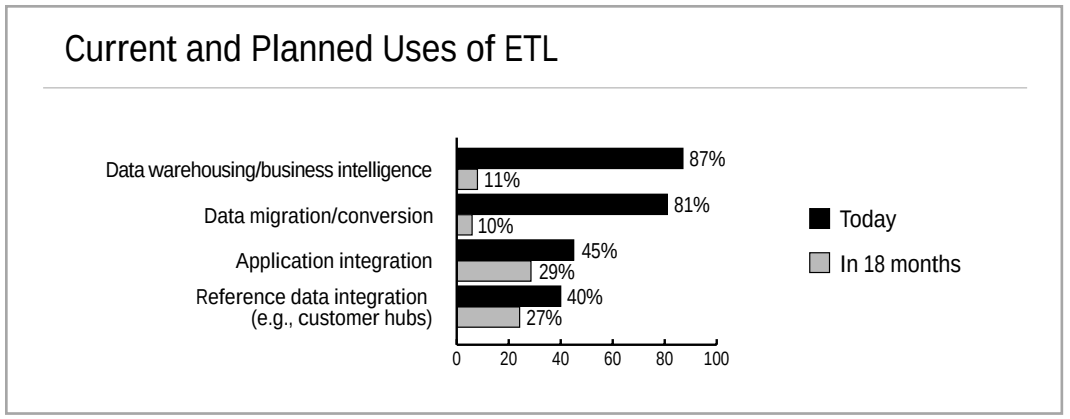


Illustration 6. Besides BI, organizations want to use ETL products for a variety of data integration needs. Based on 740 responses.

SUMMARY

As data warehouses amass more data from more sources, BI teams need ETL tools to process data more quickly, efficiently, and reliably. They also need ETL tools to assume a larger role in coordinating the flow of data among source and target systems and managing data consistency and integrity. Finally, many want ETL vendors to provide more complete solutions, either to meet an organization's enterprise infrastructure needs or deliver an end-to-end BI solution.

Build or Buy?

First Decision. As BI teams examine their data integration needs, the first decision they need to make is whether to build or buy an ETL tool. Despite the ever-improving functionality offered by vendor ETL tools, there are still many organizations that believe it is better to hand write ETL programs than use off-the-shelf ETL software. These companies point to the high cost of many ETL tools and the abundance of programmers on their staff to justify their decision.

Most Companies Buy ETL Tools

Strong Market to "Buy." Nonetheless, the vast majority of BI teams (82 percent) have purchased a vendor ETL tool, according to our survey. Overall, 45 percent use vendor tools exclusively, while 37 percent use a combination of tools and custom ETL code. Only 18 percent exclusively write ETL programs from scratch. (See Illustration 7.)

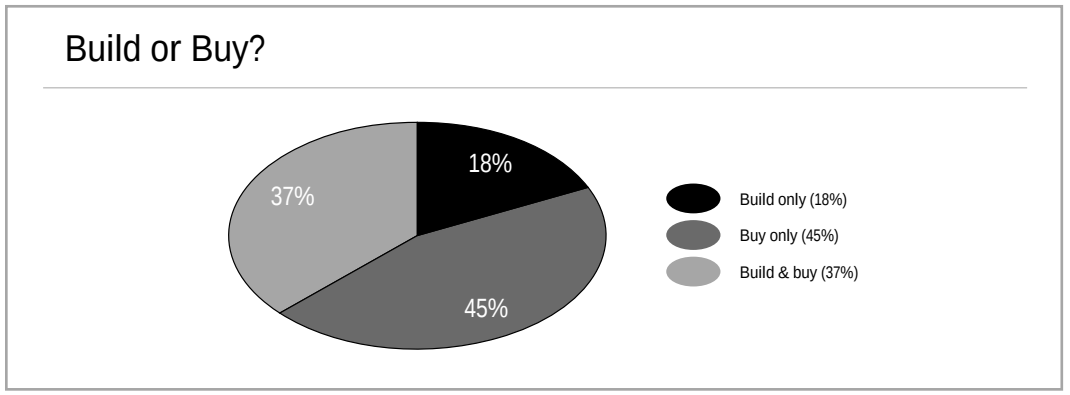


Illustration 7. The vast majority of organizations building data warehouses have purchased an ETL tool from a vendor. Based on 761 responses.

Why Buy?

Maintaining Custom Code

The primary reason that organizations purchase vendor ETL tools is to minimize the time and cost of developing and maintaining proprietary code.

"We need to expand the scale and scope of our data warehouse and we can't do it by hand-cranking code," says Richard Bowles, data warehouse developer at Safeway Stores plc in London. Safeway is currently evaluating vendor ETL tools in preparation for the expansion of its data warehouse.

Custom Code Is Laborious to Modify

The problem with Safeway's custom COBOL programs, according to Bowles, is that it takes too much time to test each program whenever a change is made to a source system or target data warehouse. First, you need to analyze which programs are affected by a change, then check business rules, rewrite the code, update the COBOL copybooks, and finally test and debug the revised programs, he says.

"Custom code equals high maintenance. We want our team focused on building new routines to expand the data warehouse, not preoccupied with maintaining existing ones," Bowles says.

Tools that Speed Deployment Save Money. Many BI teams are confident that vendor ETL tools can better help them deliver projects on time and within budget. These managers ranked "speed of deployment" as the number one reason they decided to purchase an ETL tool, according to our survey. (See Table 1.)

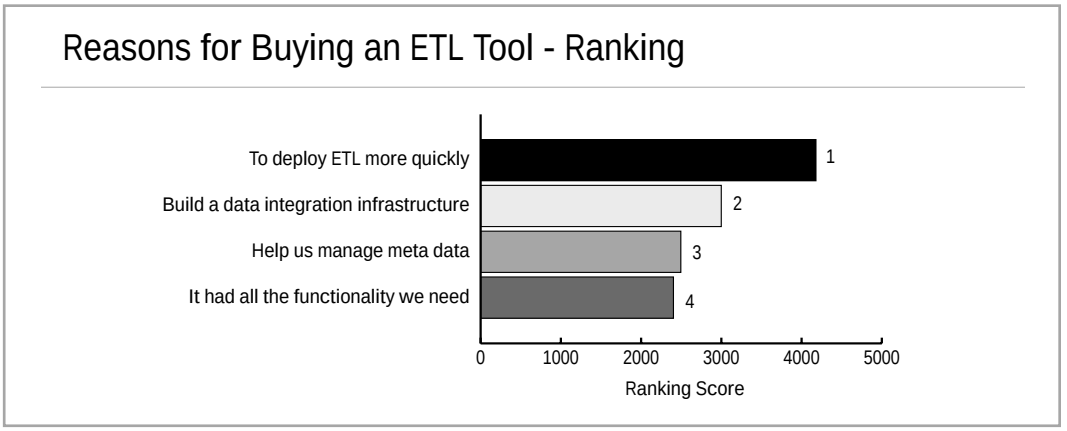


Table 1. Ranking of reasons for buying an ETL tool.

Ken Kirchner, a data warehousing manager at Werner Enterprises, Inc., in Omaha, NE, is using an ETL tool to replace a team of outside consultants that botched the firm's first attempt at building a data warehouse.

"Our consultants radically underestimated what needed to be done. It took them 15 months to code 130 stored procedures," says Kirchner. "A [vendor ETL] tool will reduce the number of people, time, and money we need to deliver our data warehouse."

EDS Canada's Light says in many situations, clients have older ETL environments where the business rules are encapsulated in COBOL or other 3GLs, making it impossible to do impact analysis. Moving to an engine-based tool allows the business rules to be maintained in a single location which facilitates maintenance."

Why Build?

Cheaper and Faster. However, some organizations believe it is cheaper and quicker to code ETL programs than use a vendor ETL tool. (See Table 2.)

"Custom software is what saves us time and money, even in maintenance, because it's geared to our business," says Mike Galbraith, director of Global Business Systems Development at Tyco Electronics in Harrisburg, PA.

The key for ETL programs—or any software development project—is to write good code, Galbraith says. "We use software engineering techniques. We write our code from specifications and a meta data model. Everything is self documenting and that has saved us a lot."

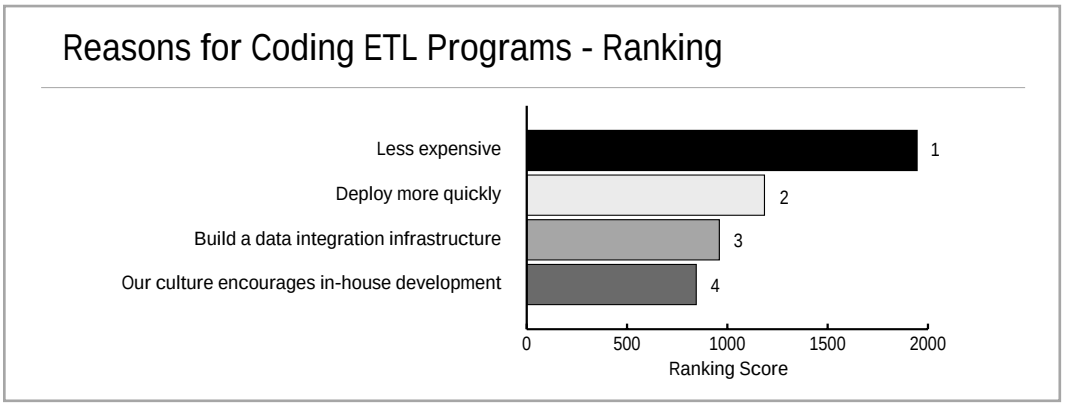


Table 2. Ranking of reasons for buying an ETL tool. Based on respondents who code ETL exclusively.

Cummins Outsources ETL Coding Overseas to Speed Development

Hubert Goodman, director of business intelligence at Cummins, Inc., a leading maker of diesel engines, says the cost to maintain his group's custom ETL code is less than the annual maintenance fee and training costs that another department is paying to use a vendor ETL tool. Goodman's team uses object-oriented techniques to write efficient, maintainable PL/SQL code. To keep costs down and speed turnaround times, he outsources ETL code development overseas to programmers in India. (Some practitioners say it is impractical to outsource GUI-based ETL development because of the interactive nature of such tools.)

The Hard Job of ETL. Several BI managers also said vendor ETL tools don't address the really challenging aspects of migrating source data into a data warehouse, specifically how to identify and clean dirty data, build interfaces to legacy systems, and deliver near-real-time data. For these managers, the math doesn't add up—why spend $200,000 or more to purchase an ETL tool that only handles the "easy" work of mapping and transforming data?

Build and Buy

Slow Migration to ETL Packages. Despite these arguments, there is a slow and steady migration away from writing custom ETL programs. More than a quarter (26 percent) of organizations that use custom code for ETL will soon replace it with a vendor ETL tool. Another 23 percent are planning to augment their custom code with a vendor ETL tool. (See Illustration 8.)

Firms Are Gradually Replacing Home Grown Code

When asked why they are switching, many BI managers said that "vendor ETL tools didn't exist" when they first implemented their data warehouse. Others said they initially couldn't afford a vendor ETL tool or didn't have time to investigate new tools or train team members.

Typically, these teams leave existing hand code in place and use the ETL package to support a new project or source system. Over time, they displace custom code with the vendor ETL

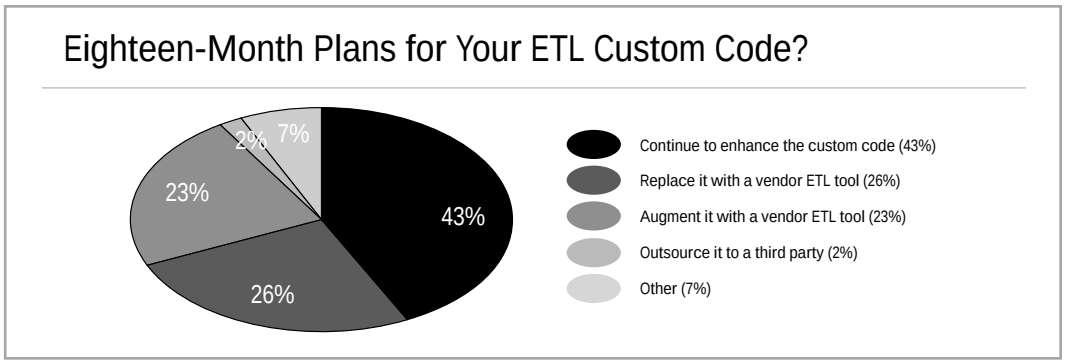


Illustration 8. Almost half (49 percent) of respondents plan to replace or augment their custom code with a vendor ETL tool. Based on 415 respondents.

tool as their team gains experience with the tool and the tool proves it can deliver adequate performance and functionality.

Strategy: Custom Code Where Appropriate

Mix and Match. Some teams, however, don't plan to abandon custom code. Their strategy is to use whatever approach makes sense in each situation. These teams often use custom code to handle processes that vendor ETL tools don't readily support out of the box.

For example, Preetam Basil, a Priceline.com data warehousing architect, points out that Priceline.com uses a vendor ETL tool for the bulk of its ETL processing, but it has developed UNIX and SQL code to extract and scrub data from its voluminous Web logs.

User Satisfaction with ETL Tools

Users Are "Mostly Satisfied" with ETL Tools

Overall, most BI teams are generally pleased with their ETL tools. According to our survey, 28 percent of data warehousing teams are "very satisfied" with their vendor ETL tool, and 51 percent are "mostly satisfied." Vendor ETL tools get a slightly higher satisfaction rating than hand-coded ETL programs. (See Illustration 9.)

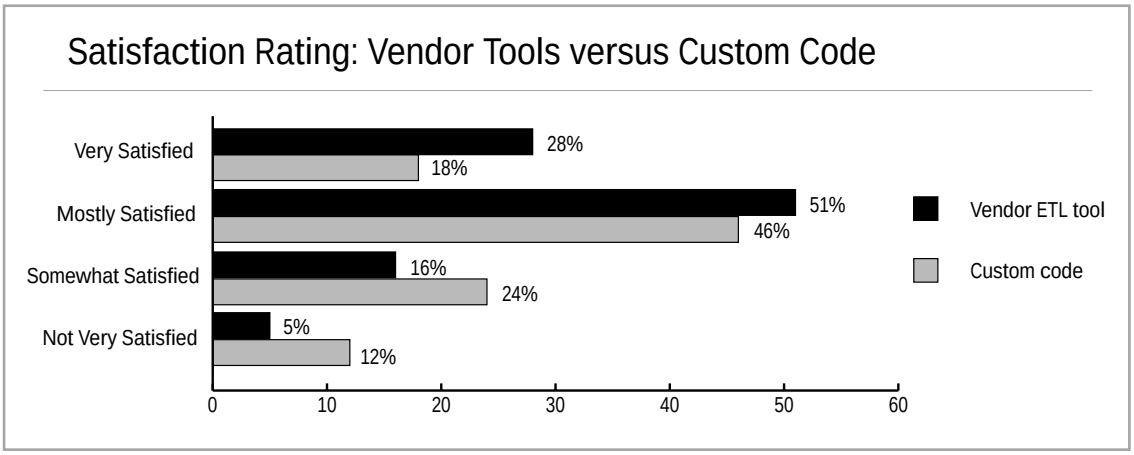


Illustration 9. Vendor ETL tools have a higher satisfaction rating than custom ETL programs for organizations that use either vendor ETL tools or custom code exclusively. Based on 341 and 134 responses, respectively.

A majority of organizations (52 percent) plan to upgrade to the next version of their ETL tool within 18 months, a sign that the team is committed to the product. Another 22 percent will maintain the product "as is" during the period. A small fraction plan to replace their vendor ETL tool with another vendor ETL tool (7 percent) or custom code (2 percent). (See Illustration 10.)

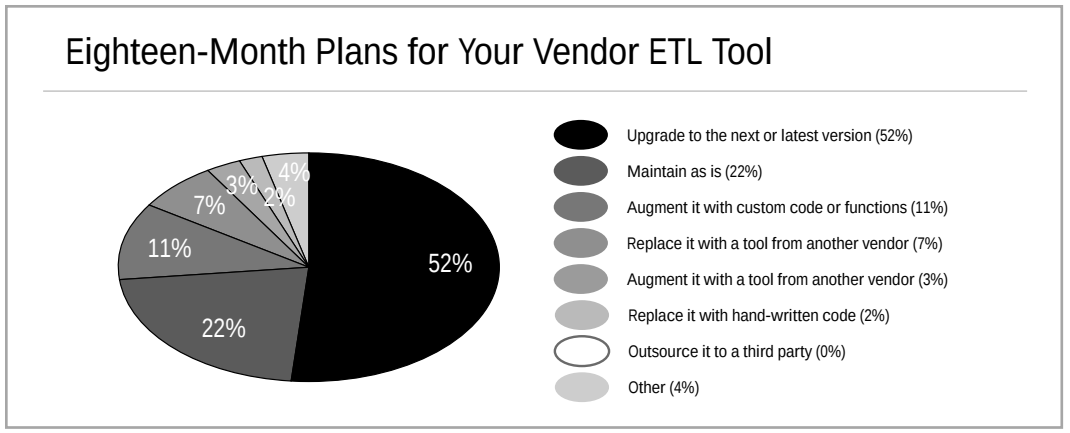


Illustration 10. Most teams are committed to their current vendor ETL tool. Based on 622 responses.

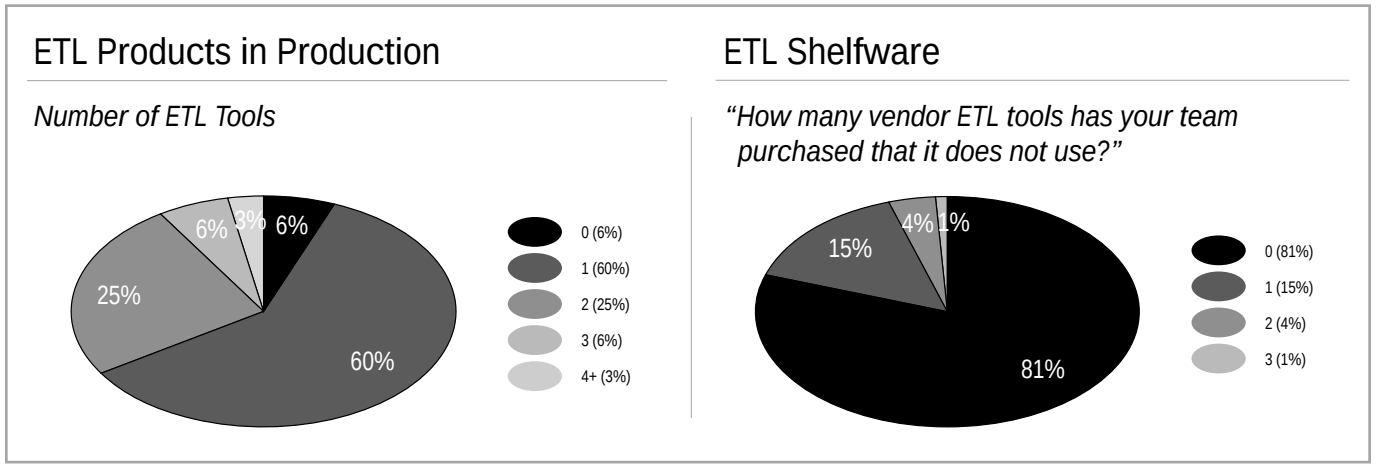


Illustration 11a. Most BI teams only use one ETL tool. Based on 617 responses.

Illustration 11b. Very few ETL tools go unused. Based on 617 responses.

Very few organizations have abandoned vendor ETL tools. Most have purchased one tool and continue to use it in a production environment. (See Illustrations 11a and 11b.) In other words, there is very little ETL shelfware.

Challenges in Deploying ETL

Data Quality Issues Top all Challenges

Despite this rosy picture, BI teams encounter numerous challenges when deploying vendor ETL tools. The top two challenges are "ensuring adequate data quality" and "understanding source data," according to our survey. (See Table 3.) Although these problems are not necessarily a function of the ETL tool or code, they do represent an opportunity for ETL vendors to expand the capabilities of their toolset to meet these needs.

Goal: Minimize Code Written Outside the ETL Tool

Complex Transformations. The next most significant challenge is designing complex transformations and mappings. "The goal of [an ETL] tool is to minimize the amount of code you have to write outside of the tool to handle transformations," says Alliance Capital Management's Connolly.

Unfortunately, most ETL tools fall short in this area. "We wish there was more flexibility in writing transformations," says Brian Koscinski, ETL project lead at American Eagle Outfitters in Warrendale, PA. "We often need to drop a file into a temporary area and use our own code to sort the data the way we want it before we can load it into our data warehouse."

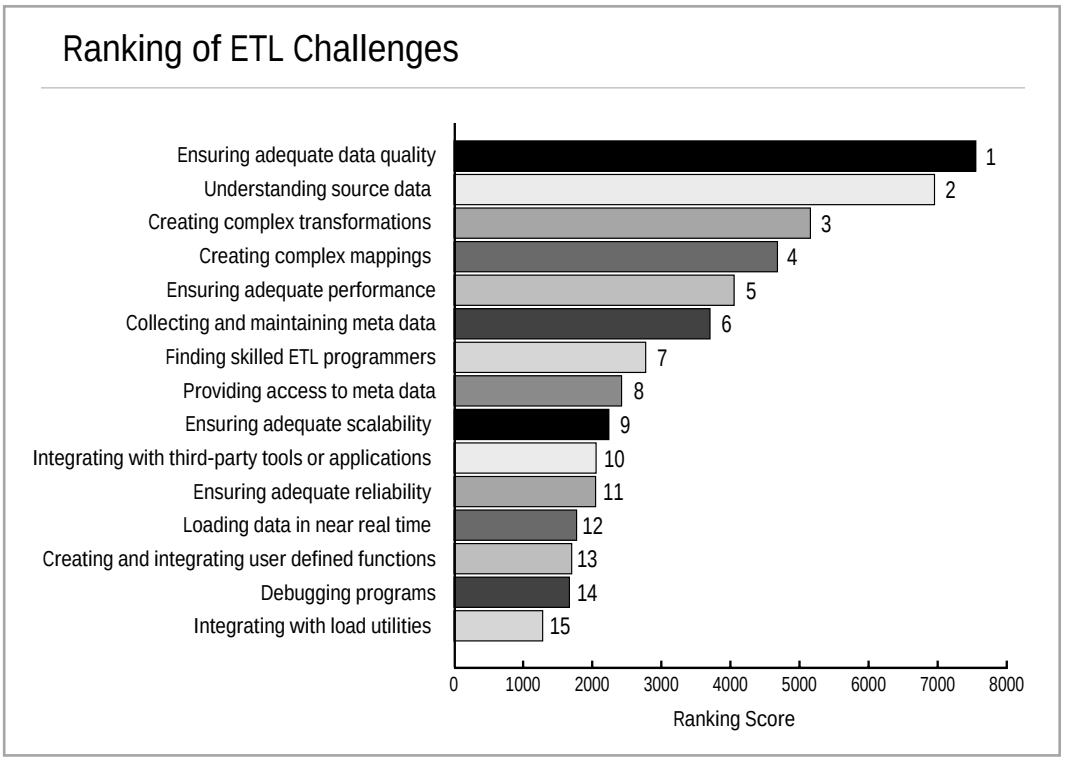


Table 3. Data quality issues are the top two challenges facing ETL developers.

According to our survey, only a small percentage of teams (11 percent) are currently extending vendor ETL tools with custom code. The rest are using a "reasonable amount" of custom code or "none at all" to augment their vendor ETL tools.

Skilled ETL Developers. Another significant challenge, articulated by many survey respondents and users interviewed for this report, is finding and training skilled ETL programmers. Many say the skills of ETL developers and familiarity with the tools can seriously affect the outcome of a project.

Priceline.com's Basil agrees. "Initially, we didn't have ETL developers who were knowledgeable about data warehousing, database design, or ETL processes. As a result, they built inefficient processes that took forever to load."

EDS Canada's Light says it's imperative to hire experienced ETL developers at the outset of a project, no matter what the cost, to establish development standards. Otherwise, the "problems encountered down the line will be much larger and harder to grapple with." Others add that it's wise to hire at least two ETL developers in case one leaves in midstream.

Developers Take 12 or More Months to Be Proficient with an ETL Tool

Most BI managers are frustrated by the time it takes developers to learn a new tool (about three months) and then become proficient with it (12 or more months.) Although the skill and experience of a developer counts heavily here, a good ETL development environment—along with high-quality training and support offered by the vendor—can accelerate the learning curve.

The Temptation to Code. Developers often complain that it takes much longer to learn the ETL tool and manipulate transformation objects than simply coding the transformations from scratch. And many give in to temptation. But perseverance pays off. Developers are five to six times more productive once they learn an ETL tool compared to using a command-line interface, according to Pieter Mimno.

Pricing

Sticker Shock. Perhaps the biggest hurdle with vendor ETL tools is their price tag. Many data warehousing teams experience sticker shock when vendor ETL salespeople quote license fees. Many teams weigh the perceived value of a tool against its list price—which often exceeds $200,000 for many tools—and walk away from the deal.

"For $200,000 you can buy at least 4,000 man hours of development and the tool's ongoing maintenance fee pays for one-half of a head count," says data warehousing consultant Darrell Piatt.

Hard to Justify the Price Compared to Using Programmers

The price of ETL tools is a "difficult pill to swallow" for companies that don't realize that a data warehouse is different from a transaction system, says Bryan LaPlante, senior consultant with Pragmatek, a Minneapolis consultancy that works with mid-market companies. These companies don't realize that an ETL tool makes it easier to keep up with the continuous flow of changes that are part and parcel of every data warehousing program, he says.

But even companies that understand the importance of ETL tools find it difficult to open their checkbooks. "Although ETL tools are much more functional than they were five years ago, most are still not worth buying at list price," says Piatt.

Spread the Wealth. To justify the steep costs, many teams sell the tool internally as an infrastructure component that can support multiple projects. "The initial purchase is expensive, but it is easier to justify when you use it for multiple projects," says Heath Hatchett, group manager of business intelligence engineering at Intuit in Mountain View, CA. Others negotiate with vendors until they get the price they want. "It's a buyers' market right now," says Safeway's Bowles.

Downward Pricing Pressures Now Exist

Given the current economy and the increasing maturity of the ETL market, the list prices for ETL products are starting to fall. The downward pressure on prices will continue for the next few years as software giants bundle robust ETL programs into core data management programs, sometimes at no extra cost, and BI vendors package ETL tools inside of comprehensive BI suites. In addition, a few start-ups now offer small-footprint products at extremely affordable prices.

Recommendations

Every BI project team has different levels and types of resources to meet different business requirements. Thus, there is no right or wrong decision about whether to build or buy ETL capabilities. But here are some general recommendations to guide your decision based on our research.

Buy If You …

- **Want to Minimize Hand Coding.** ETL tools enable developers to create ETL workflows and objects using a graphical workbench that minimizes hand coding and the problems that it can create.
- **Want to Speed Project Deployment.** In the hands of experienced ETL programmers, an ETL tool can boost productivity and deployment time significantly. (However, be careful, these productivity gains don't happen until ETL programmers have worked with a tool for at least a year.)
- **Want to Maintain ETL Rules Transparently.** Rather than bury ETL rules within code, ETL tools store target-to-source mappings and transformations in a meta data repository accessible via the tool's GUI and/or an API. This reduces downstream maintenance costs and insulates the team when key ETL programmers leave.
- **Have Unique Requirements the Tool Supports.** If the tool supports your unique requirements—such as providing adapters to a packaged application your company just purchased—it may be worthwhile to invest in a vendor ETL tool.

- **Need to Standardize Your BI Architecture.** Purchasing a scalable ETL tool with sufficient functionality is a good way to enable your organization to establish a standard infrastructure for BI and other data management projects. Reusing the ETL tool in multiple projects helps to justify purchasing the tool.

Build If You …

- **Have Skilled Programmers Available.** Building ETL programs makes sense when your IT department has skilled programmers available who understand BI and are not bogged down delivering other projects.
- **Practice Sound Software Engineering Techniques.** The key to coding ETL is writing efficient code that is designed from a common meta model and rigorously documented.
- **Have a Solid Relationship with a Consulting Firm.** Consultancies that you've worked with successfully in the past and understand data warehousing can efficiently write and document ETL programs. (Just make sure they teach you the code before they leave!)
- **Have Unique Functionality.** If you have a significant number of complex data sources or transformations that ETL tools can't support out of the box, then coding ETL programs is a good option.
- **Have Existing Code.** If you can reuse existing, high-quality code in a new project, then it may be wiser to continue to write code.

Data Integration Platforms

As mentioned earlier, many vendors are extending their ETL products to meet new user requirements. The result is a new generation of ETL products that TDWI calls *data integration platforms*. These products extend ETL tools with a variety of new capabilities, including data cleansing, data profiling, advanced data capture, incremental updates, and a host of new source and target systems. (See Illustration S3.)

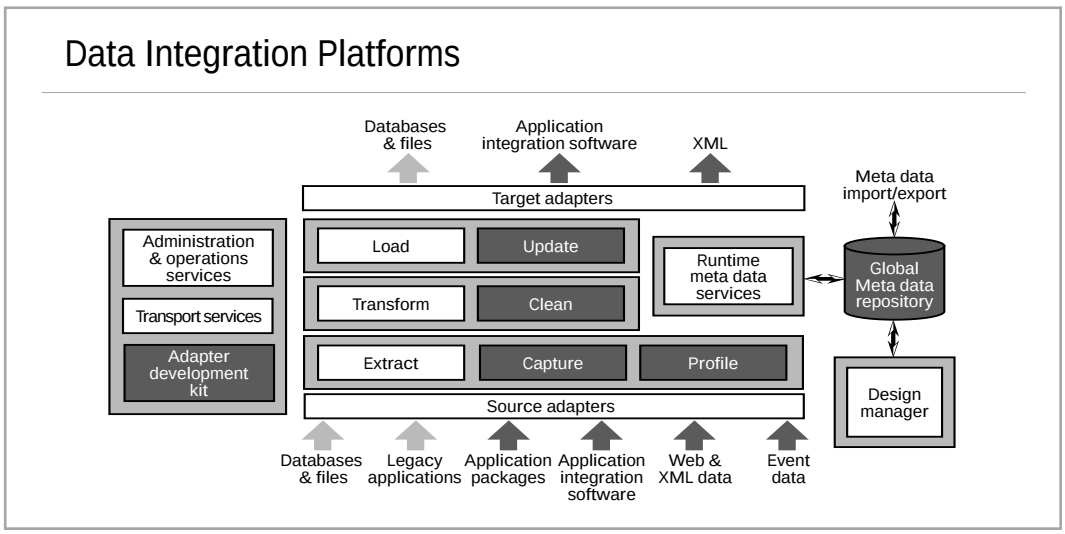


Illustration S3. Data integration platforms extend the capabilities of ETL tools.

We can divide the main features of a data integration platform into platform characteristics and data integration characteristics. No ETL product today delivers all the features outlined below.

However, most ETL vendors are moving quickly to deliver full-fledged data integration platforms.

Platform Characteristics:

- High performance and scalability
- Built-in data cleansing and profiling
- Complex, reusable transformations
- Reliable operations and robust administration

Data Integration Characteristics:

- Diverse source and target systems
- Update and capture facilities
- Near-real-time processing
- Global meta data management

High Performance and Scalability

Parallel Engines and High-Performance Computing Platforms

Parallelization. Data integration platforms offer exceedingly high throughput and scalability due to parallel processing facilities that exploit high-performance computing platforms.

The parallel engine processes workflow streams in parallel using multiple operating system threads, multi-pass SQL, and an in-memory data cache to hold temporary data without storing it to disk. Where dependencies exist among or within streams, the engine uses pipelining to optimize throughput.

The high-performance computing platform supports clusters of servers running multiple CPUs and load balancing software to optimize performance under large, sustained loads.

Today's ETL Engines Can Process 150GB/Hour

Benchmarks. In the mid-1990s, ETL engines achieved a maximum throughput of 10 to 15 gigabytes of data per hour, which was sufficient to meet the needs of most BI projects at the time. Today, however, ETL engines now boast throughput of between 100 and 150 gigabytes an hour or more, thanks to steady improvements in memory management, parallelization, distributed processing, and high-performance servers.[1]

Built-In Data Cleansing and Profiling

To avoid the "code, load, and explode" syndrome mentioned earlier, good data integration platforms provide built-in support for data cleansing and profiling functionality.

Profiling Tools: the Rosetta Stone to Source Systems

Data Profiling Tools. Data profiling tools provide an exhaustive inventory of source data. (Although many companies use SQL to sample source data value, this is akin to sighting the tip of an iceberg.) Data profiling tools identify the range of values and formats of every field as well as all column, row, and table dependencies. The tool spits out reports that serve as the Rosetta Stone of your source files, helping you translate between hopelessly out-of-date copy books and what's really in your source systems.

Data Cleansing Tools. Data cleansing tools validate and correct business rules in source data. Most data cleansing tools apply specialized functions for scrubbing name and address data: (1) parsing and standardizing the format of names and addresses, (2) verifying those names/addresses against a third-party database, (3) matching names to remove duplicates, and (4) householding names to consolidate mailings to a single address.

[1] There are many variables that affect throughput—the complexity of transformations, the number of sources that must be integrated, the complexity of target data models, the capabilities of the target database, and so on. Your throughput rate may vary considerably from these rates.

Some data cleansing tools now can also apply many of the above functions to specific types of non-name-and-address data, such as products.

Data integration platforms integrate data cleansing routines within a visual ETL workflow. (See Illustration S4.) At runtime, the ETL tool issues calls to the data cleansing tool to perform its validation, scrubbing, or matching routines.

Integrating Cleansing and ETL

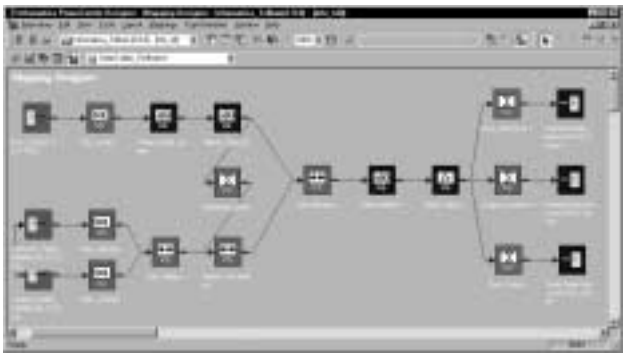


Illustration S4. Data integration platforms integrate data cleansing routines into an ETL graphical workflow.

Ideally, data integration platforms exchange meta data with profiling and cleansing tools. For instance, a data profiling tool can share errors and formats with an ETL tool so developers can design mappings and transformations to correct the dirty data. Also, a data cleansing tool can share information about matched or householded customer records so the ETL tool can combine them into a single record before loading the data warehouse.

Complex, Reusable Transformations

Extensible, Reusable Objects. Data integration platforms provide a robust set of transformation objects that developers can extend and combine to create new objects and reuse them in various ETL workflows. Ideally, developers should be able to deploy a custom object that dynamically adapts to its environment, minimizing the time developers spend updating multiple instances of the object.

Most "Reusable" ETL Objects Are Context Sensitive

"Most ETL tools today limit reuse because they are context sensitive," says Steve Tracy at Hartford Life Insurance in Simsbury, CT. "They let you copy and paste an object to support 37 instance of a source system, but you have to configure each object separately for each. This becomes painful if there is a change in the source system and you have to reconfigure all 37 instances of the object."

External Code. Although ETL tools have gained a raft of built-in and extensible transformation objects, developers always run into situations where they need to write custom code to support complex or unique transformations.

Calls to External Code Should Be Transparent

To address this pain point, data integration platforms provide internal scripting languages that enable developers to write custom code without leaving the tool. But if a developer is more comfortable coding in C, C++, or another language, the tool should also be able to call external routines from within an ETL workflow and manage those routines as if they were internal objects. In the same way, the tool should be able to call third-party schedulers, security systems, meta data repositories, and analytical tools.

Reliable Operations and Robust Administration

Users Want Greater Reliability above All Else

No matter how much horsepower a system possesses, if its runtime environment is unstable, overall performance suffers. In our survey, 86 percent of respondents rated "reliability" as a "very important" ETL feature. (See Illustration 13) This was the highest rating of any ETL feature respondents were asked to evaluate.

Robust Schedulers. Data integration platforms provide much more robust operational and administrative features than the current generation of ETL tools. For example, they provide robust schedulers that orchestrate the flow of inter-dependent jobs (both ETL and non-ETL) and run in a lights-out environment. Or they interface with a third-party scheduling tool so the organization can manage both ETL and non-ETL jobs within the same environment in a highly granular fashion.

Conditional Execution. Job processing is more reliable because data integration platforms are "smart" enough to perform conditional execution based on thresholds, content, or inter-process dependencies. This reduces the amount of coding needed to prevent such outages, especially as BI environments become more complex.

Debugging and Error Recovery. In addition, data integration platforms provide robust debugging and error recovery capabilities that minimize how much code developers have to write to recover from errors in the design or runtime environments. These tools also deliver reports and diagnostics that are easy to understand and actionable. Instead of logging what happened and when, users want the tools to say *why* it happened and *recommend* fixes.

Enterprise Consoles. Surprisingly, only 18 percent of companies said a single ETL operations console and graphical monitoring was "very important." (See Illustration 15) This indicates that most organizations do not yet need to manage multiple ETL servers. This will change as ETL usage grows and organizations decide to standardize the scheduling and management of ETL processes across the enterprise.

We suspect most organizations will use in-house enterprise systems management tools, such as Computer Associates' Unicenter, to manage distributed ETL systems.

Diverse Source and Target Systems

Intelligent Adapters. The mark of a good data integration platform is the number and diversity of source and target systems it supports. Besides supporting relational databases and flat files—which most ETL tools support today—data integration platforms provide adapters to intelligently connect to a variety of complex and unique data stores.

Application Interfaces and Data Types. For example, operational application packages, such as SAP's R/3, contain unique data types, such as pooled and clustered tables, as well as several different application and data interfaces. Data integration platforms offer native support for these interfaces and can intelligently handle unique data types.

Web Services and XML Are Fundamental Services for ETL

Web Services and XML. Data integration platforms also provide ample support to the Web world. We expect XML and Web Services to constitute an ever greater portion of the processing that ETL tools perform. XML is becoming a standard means of interchanging both data and meta data. Web Services (which use XML) is becoming the standard method for enabling heterogeneous applications to interact over the Internet. Not surprising, almost half (47 percent) of survey respondents said Web Services were an "important" or "fairly important" feature for ETL tools to possess. It's likely that Web Services and XML will become a *de facto* method for

streaming XML data from operational systems to ETL tools and from data warehouses to reporting and analytical tools and applications.

A Remedy to Spreadmarts?

Desktop Data Stores. Also, data integration platforms support pervasive desktop applications, such as Microsoft Excel and Access. By connecting to these sources via XML and Web Services, data integration platforms enable organizations to halt the growth of "spreadmarts" and the organizational infighting they create.

Adapter Development Kits. Finally, data integration platforms provide adapter development kits that let developers create custom adapters to link to unique or uncommon data sources or applications.

Update and Capture Utilities

To meet business requirements for more timely data and negotiate shrinking batch windows, data integration platforms support a variety of data capture and loading techniques.

Incremental Updates. First data integration platforms update data warehouses incrementally instead of rebuilding or "refreshing" them (i.e., dimensional tables) from scratch every time. More than one-third of our survey respondents (34 percent) said incremental update is a "very important" feature in ETL tools. (See Illustration 14.)

Full Refresh versus Incremental Updating

Change Data Capture. To incrementally update a data warehouse, ETL tools need to detect and download only the changes that occurred in the source system since the last load, a process called "change data capture." (Sometimes, the source system has a facility that identifies changes since the last extract and feeds them to the ETL tool.) The ETL tool then uses SQL update, insert, and delete commands to update the target file with the changed data.

This process is more efficient than "bulk refreshes" where the ETL tool completely rebuilds all dimension tables from scratch and appends new transactions to the fact table in a data warehouse. More than a third (38 percent) of our respondents said that change data capture was a very important feature in ETL tools. (See Illustration 14.)

Organizations Are Loading Data More Frequently

More Frequent Loads. Another way ETL tools can minimize load times is to extract and load data at more frequent intervals—every hour or less—executing many small batch runs throughout the day instead of a single large one at night or on the weekends.

For example, Intuit Inc. uses a vendor ETL tool to continuously extract data every couple of minutes from an Oracle 11i source system and load it into the data warehouse, according to Intuit's Hatchett. However, for its large nightly loads, it uses hardware clusters on its ETL server and load balancing to meet its ever-shrinking batch windows.

One way to accommodate continuous loads without interfering with user queries, is for administrators to create mirror images of tables or data partitions. The ETL tool then loads the mirror image in the background while end users query the "active" table or partition in the foreground. Once the load is completed, administrators switch pointers from the original partition to the newly updated one.[2]

Bulk Loading. Data integration platforms leverage native bulk load utilities to "block slam" data into target databases rather than simply using SQL inserts and updates. Data integration platforms can update target tables by selectively applying SQL within a bulk load operation, something that tools today can't support.

[2] Administrators must make sure that the ETL tool or database does not update summary tables in the target database until a "quiet period"—usually at night—and users must be educated about data's volatility during the period when it's being continuously updated.

Trickle Feeding Data in Near Real Time

Near-real-time Data Capture. Finally, organizations seek ETL tools that capture and load data into the data warehouse in near real time. This "trickle feed" process is needed because business users want integrated data delivered on a timelier basis (i.e., the previous day, hour, or minute—so they can make critical tactical and operational decisions without delay).

Priceline.com, for example, uses middleware to capture source data changes into a staging area and then use a vendor ETL tool to load it into the data warehouse. It has built a custom program to feed multiple sessions in parallel to its vendor ETL tool, but is planning to upgrade a new version that supports parallel processing soon.

EAI Software. As mentioned earlier in this report, most ETL tools today partner with EAI tools to deliver near-real-time capabilities. The EAI tools typically use application-level interfaces to peel off new transactions or events as soon as they are generated by the source application. They then deliver these events to any application that needs them either immediately (near real time), at predefined intervals (scheduled), or when the target application asks for them (publish and subscribe).

Data Integration Platforms Merge the Best of ETL and EAI

Bi-directional, Peer-to-Peer Interfaces. Although EAI tools are associated with delivering real-time data, they support a variety of delivery methods. Also, in contrast to ETL tools, their application interfaces are typically bidirectional and peer-to-peer in nature, instead of unidirectional and hub-and-spoke. EAI software flows data and events back and forth among applications, each alternatively serving as sources or targets.

Unlike most of today's ETL tools, data integration platforms will blend the best of both ETL and EAI capabilities in a single toolset. Few organizations want to support dual data integration platforms and would prefer to source ETL and EAI capabilities from a single vendor.

The upshot is that data integration platforms will process data in batch or near real time depending on business requirements. They will also support unidirectional and bidirectional flows of data and events among disparate applications and data stores using a hub-and-spoke or peer-to-peer architecture.

Global Meta Data Management

Data Integration Platforms Synchronize Meta Data

To optimize and coordinate the flow of data among applications and data stores, data integration platforms offer global meta data management services. The tools automatically document, exchange, and synchronize meta data among various applications and data stores in a BI environment.

Meta Data Repository and Interchange. To do this, the tools maintain a global repository of meta data. The repository stores meta data about ETL mappings and transformations. It may also contain relevant meta data from upstream or downstream systems, such as data modeling and analytical tools.

Common Warehouse Metamodel. More importantly, the tools exchange and synchronize meta data with other tools via a robust meta data interchange interface, such as the Object Management Group's Common Warehouse Metamodel (CWM), which is based on XML.

Openness and Meta Data Interfaces Are Important Features

According to our survey, 48 percent of respondents rated "openness" and 41 percent rated "meta data interface" as "very important" design features. (See Illustration 12) Although some ETL vendors recently announced support for CWM and although the specifications are considered robust, it's still too early to tell whether these standards will gain substantial momentum to facilitate widespread meta data integration.

Impact Analysis Reports. Data integration platforms also generate impact analysis reports that analyze the dependencies among elements used by all systems in a BI environment. The reports identify objects that may be affected by a change, such as adding a column, revising a calculation, or changing a data type or attribute.

Impact Analysis Reports Save Time and Money

Such reports "reduce the amount of time and effort required to [manage changes] and reduce the need for custom work," wrote one survey respondent. Moreover, when appropriate, the tools automatically update each other with such changes without IT intervention. (See Illustration S5 for a sample impact analysis report.) Thirty-three percent of respondents said impact analysis reports are a "very important" feature.

Impact Analysis Report

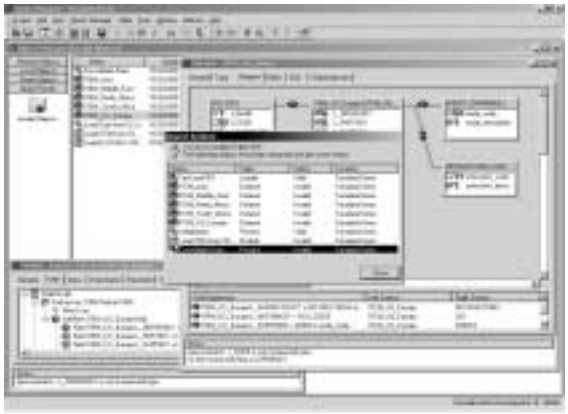


Illustration S5. ETL tools should quickly generate reports that clearly outline dependencies among data elements in a BI environment and highlight the impact of changes in source systems or downstream applications.

Data Lineage. In addition, data integration platforms offer "data lineage" reports that describe or depict the origins of a data element or report and the business rules and transformations that were applied to it as it moved from a source system to the data warehouse or a data mart. Data lineage reports help business users understand the nature of the data they are analyzing.

SUMMARY

Data integration platforms represent the next generation of ETL tools. These tools extend current ETL capabilities with enhanced performance, transformation power, reliability, administration, and meta data integration. The platforms also support a wider array of source and target systems and provide utilities to capture and load data more quickly and efficiently, including supporting near-real-time data capture.

ETL Evaluation Criteria

Guidelines for Selecting ETL Products. It may take several years for major ETL tools to evolve into data integration platforms and support all the features described above. In the meantime, you may need to select an ETL tool or reevaluate the one you are using. If so, there are many things to consider. The following list is a detailed—but by no means comprehensive—list of evaluation criteria.

It's important to remember that not all the criteria below may be applicable or important to your environment. The key is to first understand your needs and requirements and then build criteria from there. The list that follows can help trigger ideas for important features that you may need.

Focus on Your Requirements

We've tried to provide descriptive attributes to qualify the requirements so you can understand the *extent* to which a tool meets the criteria. If asked, all vendors will say their tool can perform certain functions. It's better to phrase questions in a way that lets you ascertain their true level of support.

Available Resources

ETL Matrix. The evaluation list below is a summary version of a matrix document that TDWI Members can access at the following URL: www.dw-institute.com/etlmatrix/. Another set of criteria is available at: www.clickstreamdatawarehousing.com/ClickstreamETLCriteria.htm.

ETL Product Guide. If you would like to view a complete list of ETL products, go to TDWI's *Marketplace Online* (www.dw-institute.com/marketplace) and click on the "Data Integration" category, and then select the "DW Mapping and Transformation" subcategory, and related subcategories.

Courses on ETL. Finally, TDWI also offers several courses on ETL, including TDWI *Data Acquisition, TDWI Data Cleansing, and Evaluating* and *Selecting ETL and Data Cleansing Tools.* See our latest conference and seminar catalogs for more details (www.dw-institute.com/education/).

Vendor Attributes

How Stable Is the Vendor?

Before examining tools, you need to evaluate the company selling the ETL tool. You want to find out (1) whether the vendor is financially stable and won't disappear, and (2) how committed the vendor is to its ETL toolset and the ETL market. Asking these questions can quickly shorten your list of vendors and the time spent analyzing products!

For more details, ask questions in the following areas:

- Mission and Focus
  - o Primary focus of company
  - o Percent of revenues and profits from ETL
  - o ETL market and product strategy
- Maturity and Scale
  - o Years in business. Public or private?
  - o Number of employees? Salespeople? Distributors? VARS?
- Financial Health
  - o Y-Y trends in revenues, profits, stock price
  - o Cash reserves; Debt
- Financial Indicators
  - o License-to-service revenues
  - o Percent of revenues spent on R&D
- Relationships
  - o Number of distributors, OEMs, VARs

Overall Product Considerations

Before diving in and evaluating an ETL tool's features and functions, make sure you have the right mindset. The first principle of tool selection is to question and verify everything. "It's never as easy as the vendor says," writes a BI professional from a major electronics retailer. "Verify everything they say down to the last turn of the screw before the sale is complete and the software license agreement is signed."

Platforms. After adopting a skeptical attitude, step back and see if the product meets your overall needs in key areas. Topping the list is whether the product runs on platforms that you have in house and will work with your existing tools and utilities (e.g., security, schedulers, analytical tools, data modeling tools, source systems, etc.).

You Don't Want a Partial Solution

Full Lifecycle Support. Second, find out whether the product supports all ETL processes and provides full lifecycle support for the development, testing, and production of ETL programs. You don't want a partial solution. Also, make sure it supports the preferred development style of your ETL developers, whether that's procedural, object-oriented, or GUI-driven.

Requisite Performance and Functionality. Finally, make sure the tool provides adequate performance and supports the transformations you need. The only way to validate this is to have vendors perform a proof of concept (POC) with your data. Although a POC may take several days to set up and run, it is the only way to really determine whether the tool will meet your business and technical requirements and is worth buying. You should also never pay the vendor to perform a POC or agree to purchase the product if they complete the POC successfully.

"It's really important to take these tools for a test drive with your data," says EDS Canada's Light. Mimno adds, "A proof of concept exposes how different the tools really are—this is something you'll never ascertain from evaluations on paper only."

Design Features

Ease of Use. Most BI teams would like vendors to make ETL tools easier to learn and use. A graphical development environment can enhance the usability of an ETL product and accelerate the steep learning curve. Almost all survey respondents (84 percent) said a visual mapping interface, for example, was either a "very important" or "fairly important" design feature.

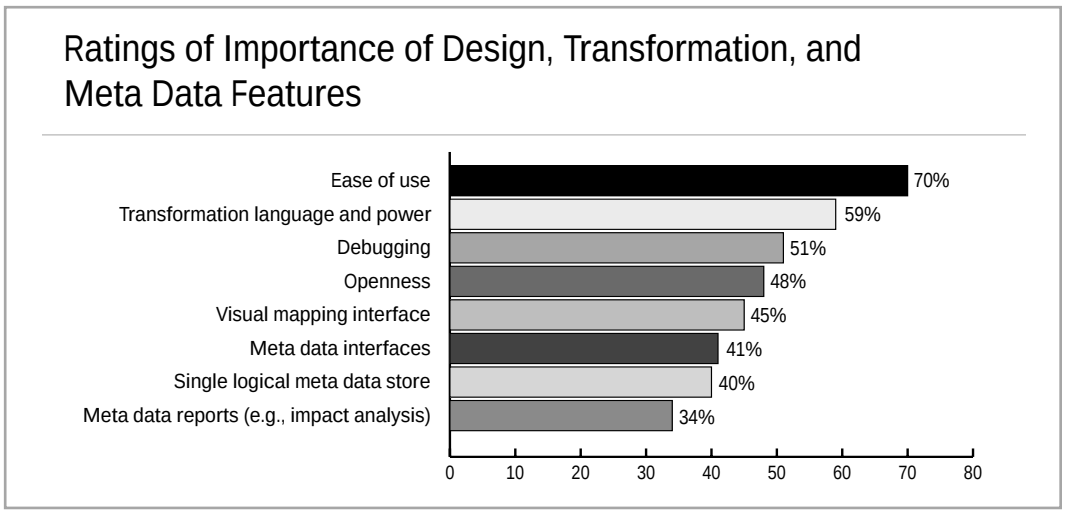


Illustration 12. Ease of use is rated as a "very important" design feature. Based on 746 respondents.

Good ETL tools save developers time by letting them combine and reuse objects, tasks, and processes in various workflows. Workflows are visual and interactive, letting developers focus on one step at a time. Wizards and cut-and-paste features also speed development, as well as interactive debuggers that don't require developers to write code.

More specifically, ask whether the ETL tool provides:

- An integrated, graphical development environment
- Interactive workflow diagrams to depict complex flows of jobs, processes, tasks, and data
- A graphical drag-and-drop interface for mapping source and target data
- The ability to combine multiple jobs or tasks into a "container" object that can be visually depicted and opened up in a workflow diagram
- Reuse of containers and custom objects in multiple workflows and projects
- A procedural or object-oriented development environment
- A scripting language to write custom transformation objects and routines
- Exits to call external code or routines on the same or different platforms
- Version control with check-in and check-out
- An audit trail that tracks changes to every process, job, task, and object

Meta Data Management Features

As the hub of a data warehousing or data integration effort, an ETL tool is well positioned to capture and manage meta data from a multiplicity of tools. Today, most ETL tools automatically document information about their development and runtime processes in a meta data repository (e.g., relational tables or a proprietary engine) which can be accessed via query tools, a documented API, or a Web browser.

Today, the best ETL tools import and export technical and business meta data with targeted upstream and downstream systems in the BI environment. They also generate impact analysis and data lineage reports across these tools. Most should be working on ways to automate the exchange of meta data using CWM and Web Services standards and keep heterogeneous BI systems in synch.

Evaluate the Reality versus Promise of a Vendor's Meta Data Strategy

For more details on meta data, ask whether an ETL product provides:

- Self-documenting meta data for all design and runtime processes
- Support for technical, business, and operational meta data
- A rich repository for storing meta data
- A hierarchical repository that can manage and synchronize multiple local or globally distributed meta data repositories
- The ability to reverse-engineer source meta data, including flat files and spreadsheets
- A robust interface to interchange meta data with third-party tools
- Robust impact analysis reports
- Data lineage reports that show the origin and derivation of an object

Transformation Features

Extensible, Reusable, Context-Independent Objects. The best ETL tools provide a rich library of transformation objects that can be extended and combined to create new, context-independent, reusable objects. The tools also provide an internal scripting language and transparent support for calls to external routines.

To learn about transformation features, ask whether the tool provides:

- A rich library of base transformation objects
- The ability to extend base objects
- Context-independent objects or containers
- Record- and set-level processing logic and transforms
- Generation of surrogate keys in memory in a single pass
- Support for multiple types of transformation and mapping functions
- The ability to create temporary files, if needed for complex transformations
- The ability to perform recursive processing or loops

Data Quality Features

The best ETL tools provide data profiling facilities, and can spawn data cleansing routines from within ETL workflows. To get details on data profiling and cleansing facilities, ask whether the ETL product:

- Provides internal or third-party data profiling and cleansing tools
- Integrates cleansing tasks into visual workflows and diagrams
- Enables profiling, cleansing and ETL tools to exchange data and meta data
- Automatically generates rules for ETL tools to build mappings that detect and fix data defects
- Profiles formats, dependencies, and values of source data files
- Parses and standardizes records and fields to a common format
- Verifies record and field contents against external reference data
- Matches and consolidates file records

Performance Features

Performance features received some of the highest ratings in our survey. (See Illustration 13.) As the volume of data increases, batch windows shrink, and users require more timely data, teams are looking to ETL tools to make up the difference.

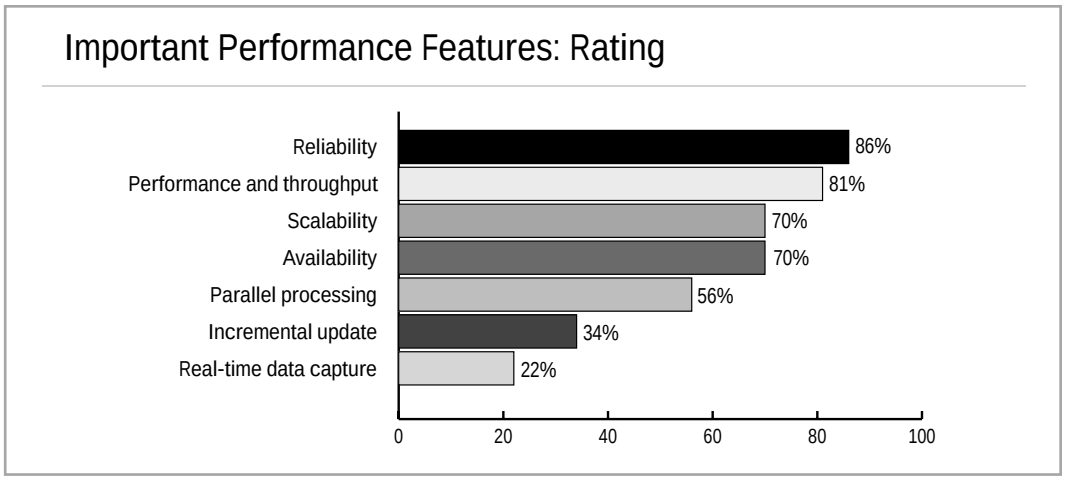


Illustration 13. Reliability and performance/throughput were rated as "very important" features by more than 80 percent of 750 survey respondents.

The best ETL tools offer a parallel processing engine that runs on a high-performance computing server. It leverages threads, clusters, load balancing, and other facilities to speed throughput and ensure adequate scalability.

To learn more about performance features, ask whether the product provides:

- A flexible, high-performance way to process and transform source data
- Intelligent management of aggregate/summary tables
  - Bundled with product or extra priced option
- High-speed processing using multiple concurrent workflows, multiple hardware processors, and systems threads and clustered servers
- In-memory cache to avoid creating intermediate temporary files
- Linear performance improvement when adding processor and memory

Extract and Capture Features

Leading ETL tools provide native support for a wide range of source databases, a top requirement among our survey respondents. (See Illustration 14.) The lowest common denominator is support for ODBC/JDBC, but make sure the tool can directly access the sources you support, preferably at no additional cost and without the use of a third-party gateway.

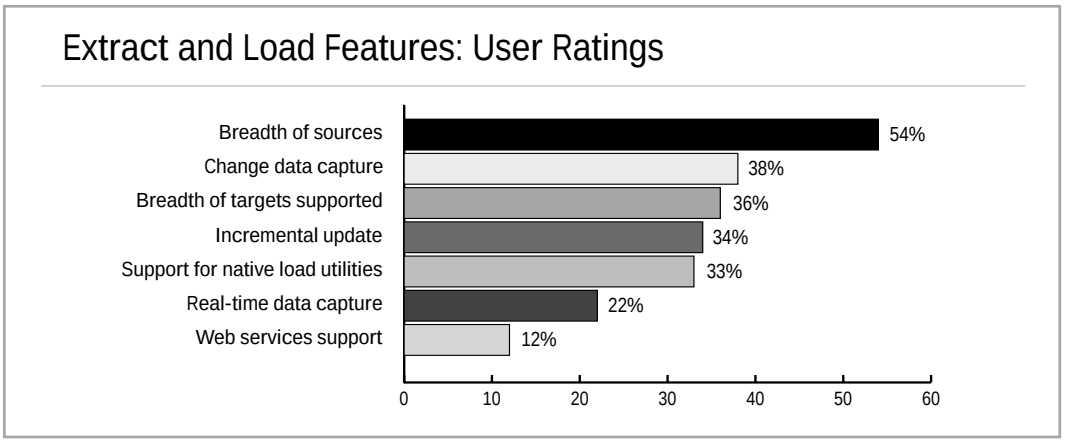


Illustration 14. Users want an ETL tool to support a wide range of sources. Based on 745 respondents who rated the above items as "very important."

Users also emphasized the importance of extracting and integrating data from multiple source systems. "The ability to use an ETL tool to extract data from three sources and combine them together is worth the effort compared to writing the routine in something like PL/SQL," says American Eagle's Koscinski.

To understand extract capabilities, ask whether the tool provides:

- The ability to schedule extracts by time, interval, or event.
- Robust adapters that extract data from various sources
- A development framework for creating custom data adapters
- A robust set of rules for selecting source data
- Selection, mapping, and merging of records from multiple source files
- A facility to capture only changes in source files.

Load and Update Features

To understand load and update features, ask whether the product can:

- Load data into multiple types of target systems
- Load data into heterogeneous target systems in parallel

- Support data partitions on the target
- Both refresh and append data in the target database
- Incrementally update target data
- Support high-speed load utilities via an API
- Turn off referential integrity and drop indexes, if desired
- Take a snapshot of data after a load for recovery purposes, if desired
- Automatically generate DDL to create tables

Operate and Administer Component

The ultimate test of an ETL product is how well it runs the programs that ETL developers design, and how well it recovers from errors. Thus, monitoring and managing the ETL runtime environment is a critical requirement. Our survey respondents rated "error reporting and recovery" especially high (79 percent) with "scheduling" not far behind (66 percent). (See Illustration 15.)

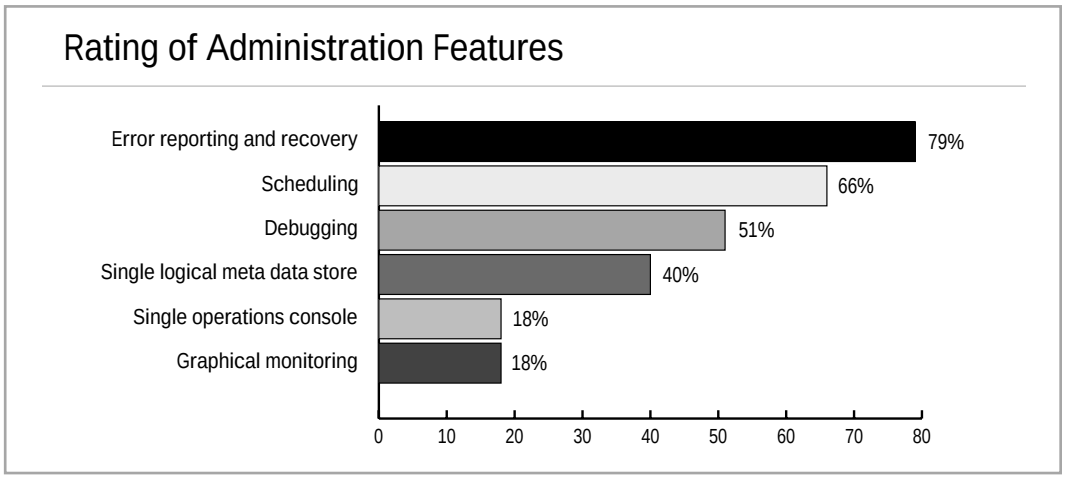


Illustration 15. Error reporting leads all administration features, according to 745 respondents who rated the above items as "very important."

For detailed information on administrative features, ask whether the product provides:

- A visual console for managing and monitoring ETL processes
- A robust, graphical scheduler to define job schedules and dependencies
- The ability to validate jobs before running them.
- A command line or application interface to control and run jobs
- Robust graphical monitoring that displays in near real time:
- The ability to restart from checkpoints without losing data
- MAPI-compliant notification of errors and job statistics
- Built-in logic for defining and managing rejected records
- Robust set of easy-to-read, useful administrative reports
- A detailed job or audit log that records all activity and events in the job:
- Rigorous security controls, including links to LDAP and other directories

Integrated Product Suites

Many vendors today bundle ETL tools within a business intelligence or data management suite. Their goal is to ensure a suite—or complete, integrated solution—will provide more value and be more attractive to more organizations.

To drill into add-on products, ask:

- Is the package geared to a specific source application or is it more generic?
- How integrated is the ETL tool with other tools in the suite?
    - o Do they have a common look and feel?
    - o Do they share meta data in an automated, synchronized fashion?
    - o Do they share a common install and management system?

Company Services and Pricing

As mentioned earlier in this report, many user organizations find the price of ETL tools too high. Almost three-quarters of survey respondents (71 percent) consider pricing among the top three factors affecting their decision to buy a vendor ETL tool. (See Illustration 16.)

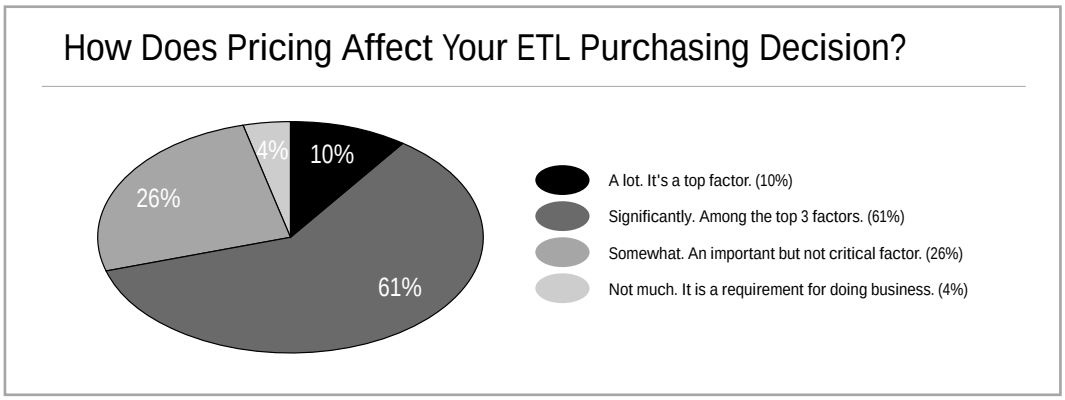


Illustration 16. Pricing is a significant factor affecting purchasing decisions. Based on 621 respondents.

Bundles versus Components. Users pay close attention to vendor packaging options. Vendors that offer all-in-one packages at high list prices can often price themselves out of contention. Vendors that break their packages into components that enable users to buy only what they need are more attractive to users who want to minimize their initial cost outlays. However, runtime charges can ruin the equation if a vendor charges the customer each time a component runs on a different platform.

More specifically, ask about:

- Pricing for a four CPU system on Windows and UNIX
- Annual maintenance fees
- Level of support, training, consulting, and documentation provided
- Pricing for add-on software and interfaces required to run the product
- Annual user conferences
- Cost of moving from a single server to multiple servers
- User group available
- Online discussion groups

Conclusion

ETL Tools Need to Keep Pace

ETL tools are the traffic cop for business intelligence applications. They control the flow of data between myriad source systems and BI applications. As BI environments expand and grow more complex, ETL tools need to change to keep pace.

Next Generation Functionality. Today, organizations need to move more data from more sources more quickly into a variety of distributed BI applications. To manage this data flow, ETL tools need to evolve from batch-oriented, single-threaded processing that extracts and loads data in bulk to continuous, parallel processes that capture data in near real time. They also need to provide enhanced administration and ensure reliable operations and high availability.

As the traffic cop of the BI environment, they need to connect to a wider variety of systems and data stores (notably XML, Web Services-based data sources, and spreadsheets) and coordinate the exchange of information among these systems via a global meta data management system. To simplify and speed the development of complex designs, the tools need a visual work environment that allows developers to create and reuse custom transformation objects.

Finally, ETL tools need to deliver a larger part of the BI solution by integrating data cleansing and profiling capabilities, EAI functionality, toolkits for building custom adapters, and possibly reporting and analysis tools and analytic applications.

It Will take ETL Vendors Several Years to Deliver Data Integration Platforms

This is a huge list of new functionality. The sum total represents a next-generation ETL product—or a data integration platform. It will take ETL vendors several years to extend their products to support these areas, but many are well on their way.

Focus on Your Business Requirements. The key is to understand your current and future ETL processing needs and then identify the product features and functions that support those needs. With this knowledge, you can then identify one or more products that can adequately meet your ETL and BI needs. ■

Business Objects
3030 Orchard Parkway
San Jose, CA 95134
408.953.6000
Web: www.businessobjects.com

Business Objects is the world's leading provider of business intelligence (BI) solutions. Business intelligence lets organizations access, analyze, and share information internally with employees and externally with customers, suppliers, and partners. It helps organizations improve operational efficiency, build profitable customer relationships, and develop differentiated product offerings.

The company's products include data integration tools, the industry's leading integrated business intelligence platform, and a suite of enterprise analytic applications. Business Objects is the first to offer a complete BI solution that is composed of best-of-breed components, giving organizations the means to deploy end-to-end BI to the enterprise, from data extraction to analytic applications.

Business Objects has more than 17,000 customers in over 80 countries. The company's stock is publicly traded under the ticker symbols NASDAQ: BOBJ and Euronext Paris (Euroclear code 12074). It is included in the SBF 120 and IT CAC 50 French stock market indexes.

DataMirror
3100 Steeles Avenue East,
Suite 1100
Markham, Ontario L3R 8T3
Canada
905.415.0310
Fax: 905.415.0340
Email: info@datamirror.com
Web: www.datamirror.com

DataMirror (Nasdaq: DMCX; TSX: DMC), a leading provider of enterprise application integration and resiliency solutions, gives companies the power to manage, monitor, and protect their corporate data in real time. DataMirror's comprehensive family of LiveBusiness™ solutions enables customers to easily and cost effectively capture, transform, and flow data throughout the enterprise. DataMirror unlocks the experience of now™ by providing the instant data access, integration, and availability companies require today across all computers in their business.

1,700 companies have gone live with DataMirror software, including Debenhams, Energis, GMAC Commercial Mortgage, the London Stock Exchange, OshKosh B'Gosh, Priority Health, Tiffany & Co., and Union Pacific Railroad.

Hummingbird Ltd.
1 Sparks Avenue
Toronto, Ontario M2H 2W1
Canada
Email: getinfo@hummingbird.com
Web: www.hummingbird.com

Headquartered in Toronto, Canada, Hummingbird Ltd. (NASDAQ: HUMC, TSE: HUM), is a global enterprise software company employing 1,300 people in nearly 40 offices around the world. Hummingbird's revolutionary Hummingbird Enterprise™, an integrated information and knowledge management solution suite, manages the entire lifecycle of information and knowledge assets. Hummingbird Enterprise creates a 360-degree view of all enterprise content, both structured and unstructured data, with a portfolio of products that are both modular and interoperable. Today, five million users rely on Hummingbird to connect, manage, access, publish, and search their enterprise content. For more information, please visit: http://www.hummingbird.com.

Informatica Corporation
Headquarters
2100 Seaport Boulevard
Redwood City, CA 94063
650.385.5000
Toll-free U.S.: 800.653.3871
Fax: 650.385.5500

Informatica offers the industry's only integrated business analytics suite, including the industry-leading enterprise data integration platform, a suite of packaged and domain-specific data warehouses, the only Internet-based business intelligence solution, and market-leading analytic applications. The Market Leading Data Integration Solution comprise of Informatica PowerCenter®, PowerMart®, and Informatica PowerConnect(tm) products, the industry's leading data integration platform helps companies fully leverage and move data from virtually any corporate system into data warehouses, operational data stores, staging areas, or other analytical environment. Offering real-time performance, scalability, and extensibility, the Informatica data integration platform can handle the challenging and unique analytic requirements of even the largest enterprises.

Membership

As the data warehousing and business intelligence field continues to evolve and develop, it is necessary for information technology professionals to connect and interact with one another. TDWI provides these professionals with the opportunity to learn from each other, network, share ideas, and respond as a collective whole to the challenges and opportunities in the data warehousing and BI industry.

Through Membership with TDWI, these professionals make positive contributions to the industry and advance their professional development. TDWI Members benefit through increased knowledge of all the hottest trends in data warehousing and BI, which makes TDWI Members some of the most valuable professionals in the industry. TDWI Members are able to avoid common pitfalls, quickly learn data warehousing and BI fundamentals, and network with peers and industry experts to give their projects and companies a competitive edge in deploying data warehousing and BI solutions.

TDWI Membership includes more than 4,000 Members who are data warehousing and information technology (IT) professionals from Fortune 1000 corporations, consulting organizations, and governments in 45 countries. Benefits to Members from TDWI include:

- *Quarterly Business Intelligence Journal*
- *Biweekly TDWI FlashPoint electronic bulletin*
- *Quarterly TDWI Member Newsletter*
- *Annual Data Warehousing Salary, Roles, and Responsibilities Report*
- *Quarterly Ten Mistakes to Avoid series*
- *TDWI Best Practices Awards summaries*
- *Semiannual What Works: Best Practices in Business Intelligence and Data Warehousing corporate case study compendium*
- *TDWI's Marketplace Online comprehensive product and service guide*
- *Annual technology poster*
- *Periodic research report summaries*
- *Special discounts on all conferences and seminars*
- *Fifteen-percent discount on all publications and merchandise*

Membership with TDWI is available to all data warehousing, BI, and IT professionals for an annual fee of $245 ($295 outside the U.S.). TDWI also offers a Corporate Membership for organizations that register 5 or more individuals as TDWI Members.

The Data Warehousing Institute
5200 Southcenter Blvd., Suite 250
Seattle, WA 98188
Local: 206.246.5059
Fax: 206.246.5952
Email: info@dw-institute.com
Web: www.dw-institute.com

tdwi
THE DATA WAREHOUSING INSTITUTE