March 19, 2002

Market Overview Update: ETL

Lou Agosta

Contributing Analysts: Keith Gile

Giga Position

As Giga predicated in May 2001, extract, transform and load (ETL) technology continues to be in transition from moving dumb data to performing intelligent information integration. Due to the diversity, incompatibility and volatility of data sources in the enterprise — enterprise resource planning (ERP), legacy systems, Web logs, Extensible Markup Language (XML), message brokers, customer relationship management (CRM) — ETL is unlikely to become an open, commodity technology solution and proprietary approaches will remain dominant for the foreseeable future. However, ETL technologies bundled with the underlying database "at no extra charge" or at a significant discount — **Microsoft** DTS, **Oracle** Warehouse Builder, **IBM** Data Warehouse Manager — are likely to consume a growing market share, especially in entry-level ETL installations. While the corporate IT function may be reluctant to *buy* its ETL technology from its database vendor, it is a different matter when that ETL technology ships at no extra charge or at minimum cost. Therefore, IT developers and administrators should leverage the ETL technology that ships with the latest versions of their relational databases. However, the IT department will continue to have to qualify and purchase more functionally rich (and more expensive) best-of-breed ETL tools as justified by client-specific information integration requirements that aim at complex and deep solutions not envisioned by the generic products.

Since last May 2001 when Giga last reported on the overall ETL market and the year ending December 2001, the size of the market grew from $614 million to $667 million, a growth rate of slightly more than 11 percent on an annual basis. This is down considerably from the 60 percent growth rate enjoyed in 2000, and significantly off from the rate of 30 percent predicted by Giga in May 2001 for last year. The was due to overall effects of the economic downturn as well as lengthening sales cycle for the capital acquisition process for six-figure software technologies. Given the increase over 2001 in inquiries received by Giga's data warehousing practice about ETL technologies and issues, interest remains strong and is still a part of mainstream infrastructure build out in the data warehousing technology stack. Therefore, Giga is expecting a return to a more robust growth rate of 15 percent if the economic recovery proves to be a sustained one.

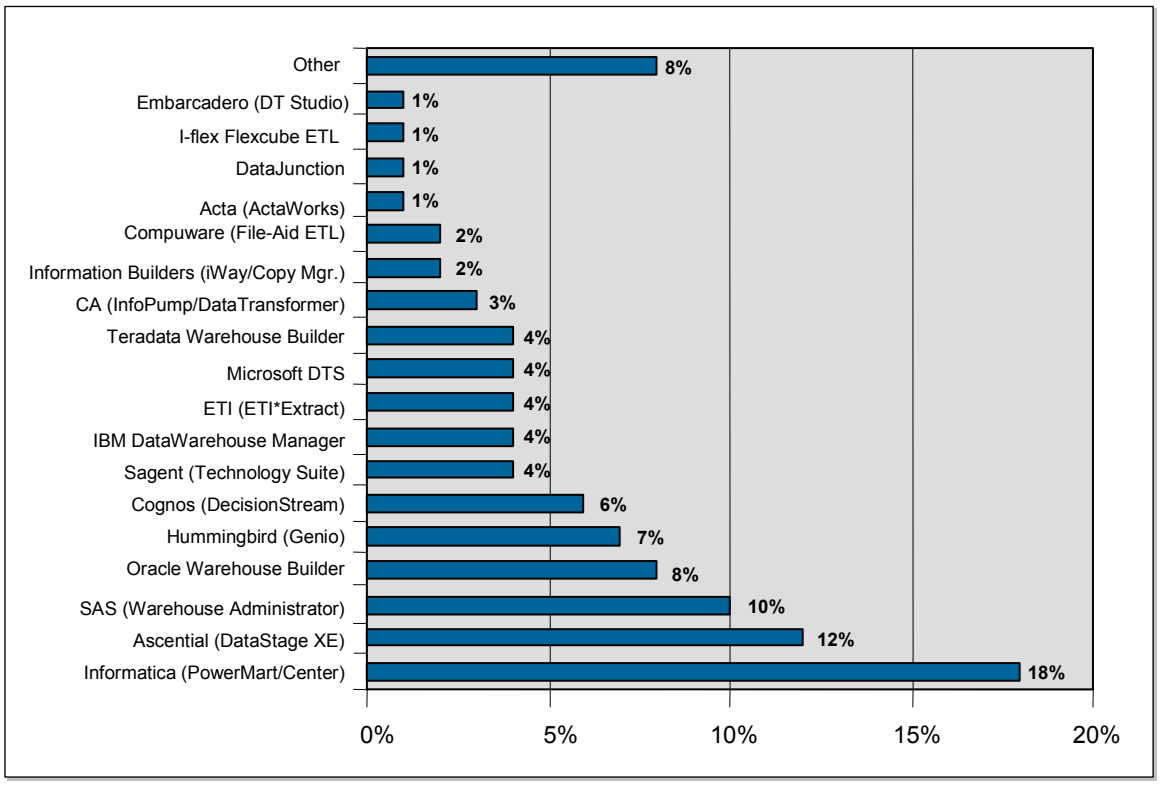
Proof/Notes

Market Dynamics

The area of the market that shows the most growth is what might be described as the entry-level segment of the market as constituted by Microsoft DTS and IBM Data Warehouse Manager. Individually the market shares have virtually doubled, admittedly from a low level, with IBM's Data Warehouse Manager going from 2 percent to 4 percent while Microsoft grew from 2 percent to 6 percent. Collectively the entry-level approach has grown from 11 percent to 16 percent of the market when Oracle Warehouse Builder (OWB) is included (though the latter is a separately licensable product). The statistical tie that existed for first place has been broken with **Informatica** pulling into the lead. Ascential's ETL revenues have also grow, reportedly at a 27 percent rate. However, relative to the overall market, a decrease occurred because ETL revenues are no longer aggregated with those of the **Informix** database, making it possible for the first time ever to obtain a more exact estimate of license revenues. **Ascential** has been distracted by its merger with Informix and subsequent sale of Informix's product line to IBM. Ascential has now used some $46 million of the $1 billion in cash with the acquisition of **Torrent** parallel technology (see IdeaByte, An Unparalleled Acquisition:

Ascential Software Buys Torrent Systems, Lou Agosta). On March 12, 2002, Ascential announced it would be acquiring data quality provider **Vality**, underscoring the importance of checking data quality at "ETL time." Both these acquisitions will serve Ascential well in staging a comeback.

Figure 1: The ETL Market — $667 Million



Source: Giga Information Group

Trends

- **ETL technology shipping "at no extra charge" with the underlying relational database**: Microsoft established this trend in 1999 when Data Transformation Services (DTS) shipped with MS SQL Server as part of the database, and it is an ongoing key driver of the market. DTS includes a relational schema, an OLAP cube, ETL (i.e., DTS) functionality and metadata interoperation. IBM followed in 2000 as the Visual Warehouse was rearchitected and rebranded as the Data Warehouse Center (DWC). Taking a lesson from Microsoft, DB2 DWC ships at no extra charge with its DB2 UDB database installations running on Windows NT. However, versions other than NT require an extra license fee and provide additional capabilities. DWC provides a nice entry ramp to Data Warehouse Manager (DWM). Some 100 built-in data transformations are provided as part of DWM. IBM now recommends Ascential DataStage instead of ETI when installations seek to move to deeper or more complex ETL functionality. Oracle's stated direction is to embed ETL (along with OLAP and data mining) in the Oracle 9i R2 database that acts as a single transformation engine (tentatively scheduled for the second quarter of 2002). The ETL processing occurs within the Oracle 9i Database, using PL/SQL and ETL-focused SQL, which is likewise the case for installations that move up to Oracle Warehouse Builder (OWB), a separately licensable product, where table functions in Java and C are also supported.
- **Integrate with ERP and CRM packages**: Different vendors use different metaphors for intelligent

information integration — adapters, connectors, interpreters — but the idea of preconfigued metadata is central to each. The target data store (usually, but not always, a data warehouse) is made intelligent by means of metadata that defines the source data model in the underlying **SAP**, **PeopleSoft**, **J.D. Edwards** ERP or **Siebel** CRM systems. Such adapters add the "intelligence" to "intelligent information integration." The provisioning of additional adapters for other ERP systems and for the big four (Oracle, PeopleSoft, SAP, Siebel) by the less well known ETL data transformers, such as **iWay Software**, **Compuware** or **CA**, is a growth industry. This approach also works with legacy transactional system; however, no way is available by which an ETL tool "knows" the customer's proprietary system, so design work is required to define the source and target data structures. No ready substitute is available for the intelligence of a human designer in the case of legacy renewal systems, though the automation of a data-modeling tool will be useful.

- **Integrate with XML**: This is especially significant for enterprises with commitment to business-to-business (B2B) e-commerce where XML is becoming the near universal language of data exchange and interoperability. XML adapters that enable the major ETL vendors to consume and produce XML out of the ETL engine have been shipping for more than a year, although the adoption is still sparse. This is because XML is such a discontinuous technology that end-user installations are still learning (and defining) the rules of the game. The incremental advance of the much hyped Web Services is likely as XML-enabled database stored procedures are connected to the e-commerce extranet (see Planning Assumption, The Application Platform Battle: Web Services Just One Dimension of a Resilient Platform Strategy, Randy Heffner and Mike Gilpin). Since XML is hierarchical, including nested occurrences of structures, developers should be sure the adapters with their XML vendor can handle deep and variable occurs structures to avoid being surprised by limited functionality. It is important to note that another separate use of XML lies "under the covers" in metadata infrastructure and exchange. This is equally important, but not directly visible to end users except to improve the interoperability of tools (and perhaps as something about which to cross-examine vendors in terms of their metadata direction).
- **Integrate with message brokers**: Message brokers are the heart of near real-time enterprise application integration (EAI). Therefore, it is important that adapters are available for all the key ETL tools — Ascential, Informatica, **Hummingbird**, SAS — to read from and write to MS Series or **Tibco** Rendezvous message queues. Informatica also has technology connectors for **Vitria** and **WebMethods**. Note that this means the ETL engine executes a record at a time. Giga cautions data and application architects that data integrity is guaranteed within the respective tools, but that maintaining the transaction across the system interface is another issue. In short, MQ Series is just another data source or target for the average ETL tool. No "two phased commit" exists across the interface. This is another proof point that the ETL and EAI technologies continue to converge, though the convergence is likely to remain partial and incomplete (see IdeaByte, Determining When ETL or EAI Technology Is Required, Lou Agosta, and the discussion below).
- **Data quality is a growth industry**: Data quality vendors such as **FirstLogic**, **DataFlux** (SAS), **Group-1**, **Trillium** (**Harte-Hanks**) and Vality, as well as data aggregators such as **Axciom**, **Experian**, **Polk** and **TransUnion**, interoperate with ETL tools to perform data standardization at ETL time. For example, Informatica and Ascential invoke data quality technology from FirstLogic. Ascential includes its own data quality software with DataStage at no extra charge and has now purchased data quality vendor Vality (expected to be available either separately or bundled as separately licensable). Also, Oracle Warehouse Builder reportedly includes data quality technology originally based on Carleton Pure Integrate. A major challenge will continue to be the building of a consolidated and unified consumer view across multiple contact points — physical retail outlets, call centers, online and new wireless gadgets. Data augmentation is on the critical path to completing the process. Data quality technology tools and algorithms will be applied to detailed transactional data to provide intelligent information integration of individual customer, household or prospect profiles at which targeted marketing can be directly. As technology catches up with the hype about the 360-degree view of the customer, designing and building a clean and consistent representation of the

customer, product or key data dimensions demonstrates that intelligent information integration is hard work (see IdeaByte, Integrated Customer Information — Data Integration Technology Catches up With the Hype, Lou Agosta). A significant challenge continues to be to leverage back-to-basics match-merge data processing functions to extract information and intelligence from the torrent of data. The IT department ought to be happy to hear from end users with these sorts of challenges — it is the sort of function IT does best (see Planning Assumption, Evaluating Data Quality Vendors: Part 1, Lou Agosta).

- **Handle large volume points by means of parallel processing**: The competition in the ETL market around managing large volume points continues to be intense. Informatica has a patent relating to parallel processing and using multiple engines of symmetric multi-processing (SMP). Genio Suite (from Hummingbird) has parallel load capabilities, which are often exploited in a Teradata (**NCR**) context. **Sagent Technology** recently reported a credible but unaudited benchmark in which its ETL technology loaded a terabyte of data in about 8.5 hours. Using the data model of customer and products that is defined by the TPC-H benchmark (see www.tpc.org), Ascential reported that it moved 1TB of data through its DataStage XE Parallel Edition in 1.4 hours. This was intended literally to benchmark processing at a base level (see IdeaByte, The Case for an ETL Benchmark, Lou Agosta). The result on a 24-processor p680 at an IBM development center required eight hours and 43 minutes to transform and load the 1TB of data. Giga disagrees with the assertion that this can be automatically scaled to run in just over three hours on a 64-way machine. Such an operation might be feasible — no one is saying it is impossible. What is being said is that it has not been proven this time (March 15, 2002). The likelihood of vendor-based "he said" vs. "she said" points toward the importance of having an audit performed by an objective third party (such as a group like the Transaction Processing Performance Council). If the hardware and its configuration are different, then we are not only testing the ETL technology, but also the entire technology stack, including the hardware. While this promises to be a source of controversy, what is *not* controversial is the growing data volumes. Giga is encouraged to see additional parallel processing capabilities being mobilized (see IdeaByte, An Unparalleled Acquisition: Ascential Software Buys Torrent Systems, Lou Agosta) to address installation data center volume requirements. See below on future prospects, since there is reason to believe the creation of an objective ETL benchmark is something whose time has come.

Future Prospects

- **Moving up market into analytic applications**: Informatica, **Acta** and **Cognos** are ETL vendors that are shipping analytic applications. In particular, these are star schema data structures, including data models and metrics, relating to decision support in such application domains as finance, marketing, inventory control, e-commerce and related business processes. In many ways, these vendors are marketing a "kinder, gentler SAP Business Warehouse," whose 300-plus InfoCubes can represent installation and operational complexities (see Planning Assumption, Market Overview: Data Warehousing Out of the Box, Lou Agosta). In what Giga considers a clever reversal, SAP is bundling Ascential DataStage and shipping an entry-level version of it with SAP Business Warehouse (BW 2.0) to make credible SAP's claim to assimilate to SAP BW nonSAP transactional data stores such as legacy systems and other ERP packages. However, in other respects this trend is nascent and the ETL vendors still garner 90 percent or more of their revenue from ETL licenses and services. Those data warehousing developers and administrators looking to add business value to data integration rapidly will soon confront the task of building, purchasing or customizing the downstream analytic application (whether it is purchased as a package or developed to accommodate a proprietary, in-house business process).
- **Convergence of ETL and EAI**: This trend has been frequently announced and long delayed. It remains a compelling vision because no IT division wants to acquire two sets of data transport technologies if a single one will do the job. This trend is both exemplified and undercut by the ability of the major ETL vendors to interface with the basic message brokers such as MQ Series or

Tibco Rendezvous. Ascential, Informatica, Hummingbird and SAS all ship adapters for at least MQ Series. (Informatica also connects with Vitria and WebMethods.) As processors become more powerful and applications more flexible, the option of moving data a record at a time (or in small bursts) in order to address near real-time information opportunities becomes less costly. Giga agrees this convergence is compelling, but believes it will remain incomplete for three reasons:

1. The varieties of heterogeneous data stores, especially back-end ERP and legacy systems, make a single technology approach untenable.
2. Metadata requirements cannot be addressed by the message broker approaches for the foreseeable future, robbing developers and administrators of the benefits of reuse and improved productivity and maintenance.
3. The fundamental distinction between query-intensive and transaction-intensive systems invites and promotes the different approaches of efficient bulk ETL vs. more rapid messaging.

ETL tools address query-intensive aggregation processes so critical to decision support the way no other solution can and are unlikely to be displaced for the foreseeable future (see IdeaByte, Determining When ETL or EAI Technology Is Required, Lou Agosta).

- **An objective, audited ETL benchmark**: Benchmarks can become a source of marketing hype and spin unless they are carefully defined and audited by independent third parties. Even then, many hazards and risks threaten the objectivity of the results. In spite of these risks, Giga believes industry standard, independently audited benchmarks such as those sponsored by the Transaction Processing Performance Council have value in driving the development of new technology features that benefit real-world customers (not just "benchmark specials"), creating an economic reference point about costs, setting a performance bar at a point in time for a given software, surfacing lessons learned about how to tune products that can be shared with the end-user community (see IdeaByte, The Case for an ETL Benchmark, Lou Agosta). Since auditing of benchmarks is a subset of objective research (which, in turn, is Giga's commitment (though we do not operate a lab)), we are interested and engaged by such an undertaking. Now may be an opportune moment to gain traction in establishing an objective ETL benchmark since business ethics — integrity in messaging, transparency in metrics (i.e., an objective benchmark), and accountability in leadership — is arguably the latest post-**Enron** business and technology trend.
- **Standards-based metadata**: Even though the standards process is never as quick as required, the Common Warehouse Metamodel (CWM) metadata specification is approved and complete enough to allow the ETL vendors to implement it. Expect ETL vendors to back into such standards relatively slowly, since doing so will undercut their own proprietary solutions that have an installed base, enjoy a measure of technology lock-in and generate a corresponding revenue stream. For that very reason XML (and the XML interchange format of CWM, XMI) will win adherents — it is a method of protecting investments already made in ETL, design and data presentation tools. Thus, XML looks like the cavalry to the rescue of metadata. The strategy is to use any metadata repository or tool that can encode and decode XMI streams to exchange metadata with other repositories or tools with the same capability. The resulting tool interoperability will mean both more end-user options and more productivity. This is the win-win cycle — that is the whole point of metadata (for a list of vendors that support CWM and related details, see IdeaByte, Reports of the Demise of Metadata Are Premature, Lou Agosta.)

Alternative View

The solution with the most vision — enterprise application integration — will advance from being the second most likely scenario to being implemented as a matter of fact. No one wants to buy two tools where only one

will be satisfactory in dong the job. Using such methods as human interface engineering, impact analysis of the relational database catalog and related design repositories, software wizards will enable the automated capture and reuse of metadata to generate "good enough," quick and easy applications to transport and transform data, thereby overcoming the many-to-many labyrinth of heterogeneous data sources. What could be more of a commodity than moving data? Relative to an ETL tool, EAI is "outside the box thinking" and as such will carry the day. In spite of the lack of historical sequence or top-down development of the technology, EAI will be perceived as a breakthrough relative to batch data movement. Zero latency data update will become an increasingly dominant part of the enterprise architecture design blueprint. Thus, a metadata solution will occur at the enterprise level, though not in sufficient time to prevent the need for rework and substantial supplementary consulting services to close the gap between automated initial implementation and hand-coded, manual maintenance. Under this scenario the transition of the ETL category from data transport to intelligent information integration is just a stop on the way to the replacement of the whole class of products by EAI or a future derivative version of EAI technology.

Findings

During 2001 the ETL market grew to some $667 million, expanding at 11 percent on an annual basis. This market continues to migrate from one that moves dumb data to one that provides intelligent information integration, analytic applications, software development and data management in the broad sense of managing information as an enterprise resource.

Key ETL market trends include:

1. Delivery of ETL technology "at no extra charge" or at a minor incremental cost with the major underlying relational databases
2. The integration of diverse enterprise data stores by the major ETL vendors to provide intelligence about customers, products and markets
3. The use of parallel processing technology to manage increasing volume points in the terabyte range
4. The availability of adaptors and connectors that enable ETL tools to interface with and handle a growing diversity of ERP and CRM packages, XML data stores and message brokers in addition to the standard relational databases and legacy systems

Future trends include the move by ETL vendors such as Informatica, Cognos and Acta up market into analytic applications (Oracle bundles its Business Intelligence Suite (BIS) with Oracle Applications), the convergence of ETL and EAI functionality (though Giga believes that the ultimate convergence will remain incomplete) and the continuing incremental advance of standards-based metadata. Other vendors — Ascential and Hummingbird — do not aspire to analytic star schemas, but rather are shipping portals as part of an intelligent information integration approach. Similarly, Oracle portlets enable Web page customization.

An objectively defined, auditable ETL benchmark is at least a year into the future, but would be a boon to both owners and prospective buyers of ETL technology, as well as the broader technology market, if vendor self-interest could be set aside or at least contained, perhaps under the auspices of the TPC as a subset of a revised TPC-H.

In general, ETL tools are more accommodating of query-intensive applications where persistent data stores are needed, whereas EAI tools are more useful in transaction-intensive environments where intelligent routing and near real-time synchronization of operational systems is critical. Once again, with the provisioning of adapters for MQ Series, the ETL vendors are addressing latency issues, though they still lack full-featured intelligent messaging.

XML is the cavalry to the rescue of metadata. The strategy is to use any metadata repository or tool that can encode and decode XMI streams to exchange metadata with other repositories or tools with the same

capability. Though the standards are reasonably well defined, this trend is largely a work-in-progress with the interoperability standards of the CWM being honored more in rhetoric than in shipping products.

Even though the standard process is never as quick as required, the Common Warehouse Metamodel metadata specification is arguably complete enough to allow vendors to implement it. They are doing so, though, in order to protest their own installed base, more slowly than most IT divisions and industry analysts might wish.

Recommendations

For IT developers and administrators, the overwhelming recommendation is to capture metadata at "ETL time," the point where data is extracted from the transactional system and loaded into the data warehouse. From an architectural perspective, the ETL tool is the logical place to capture metadata, sitting as it does astride the corporate architectural choke point where operational data can readily be inspected, transformed, aggregated and mapped into the decision-support system.

XML is the best prospect to rescue metadata, and developers and data administrators should look for ETL, design and query products that accommodate the CWM's XMI standard and with XML-interchange capabilities.

Due to the proprietary technology in the ETL tools, diligence is required in selecting a vendor, and prospective buyers should assess their information integration requirements carefully. Qualify and select an ETL with a commitment toward building a long-term relationship with "win-win" thinking in mind.

In spite of the proprietary nature of ETL tools, especially the transformation engines that tend to function as black boxes, data warehousing developers and administrators should plan on using an ETL tool where the interfaces to be developed between transactional and data warehouses are more than about seven (as a rule of thumb). In the absence of the automation provided by the ETL tool and its metadata, the development team will be on a slope of diminishing return, hand coding multiple, many-to-many interfaces between operational and decision-support systems. Therefore, IT developers and data administrators should plan on capturing and leveraging metadata by means of the ETL tool and its metadata repository as the key to productivity improvements in building and maintaining interfaces between transactional and data warehousing systems.

References

Related Giga Research

Planning Assumptions

Market Overview: ETL in Transition, Lou Agosta

The Data Warehousing ETL Tool Market Matures, Lou Agosta

Need for Analytics Spurs Growth in Business Intelligence Sector, Keith Gile

Emerging Internet Data Integration Solutions, Mike Gilpin

IdeaBytes

Extraction, Transformation and Loading in Transition, Lou Agosta

Reports of the Demise of Metadata Are Premature, Lou Agosta

ETL Testing Practices, Lou Agosta

SAS Data Warehousing Shows Strong Growth, Lou Agosta

Comparing Metadata Approaches Across EAI and ETL Tools, Lou Agosta

More CRM Without the Customer — and Without the Data Warehouse, Lou Agosta

BI Vendors Continue to Exhibit Strong Growth Despite Economic Slowdown, Keith Gile

Emerging Standards-Based Application Integration Technologies Gaining Momentum, Mike Gilpin