white paper

Data Integration and Extract, Transform, and Load (ETL) 101
Overview of Data Integration and ETL Concepts
July 2006
This paper discusses the benefits of data integration as part of an effective data-management strategy, and some of the key considerations and challenges faced by organizations that seek to maximize the return from their investment in data integration. The process of Extract, Transform, and Load (ETL) is discussed in detail, including the Profiling, Data Quality, and Integration stages.

It also discusses Enterprise Information Management (EIM), a current trend in data management that creates a synchronized platform of the key functions in information management for improved management and reliability of enterprise data.

Finally, it describes the ROI from data integration, and the customer acceptance criteria for integration tools to provide an overview of Data Integration and ETL.
Table of Contents

The Importance of Data Integration and Data Quality for Effective BI 4
Challenges of Data Integration 4
Organizational Considerations 4
Business Drivers and Challenges 5
Technical Drivers and Challenges 6
Data Integration Using Extract, Transform, and Load (ETL) 6
Stages in the ETL Process: Profile, Cleanse, and Integrate 7
The Data Profile Stage (Understanding Information) 7
The Data Quality Stage (Cleansing Information) 8
The Data Integration Stage (Transforming and Integrating Information) 8
EIM—a Comprehensive Approach to Data Integration 9
Customer Acceptance Criteria from Data Integration and ETL Tools 9
The Return on Investment (ROI) from Effective Data Integration 10
Summary 10
Sources 11
About Unisys 11
The Importance of Data Integration and Data Quality for Effective BI

Accurate and complete data is a key ingredient of effective Business Intelligence (BI). Unreliable data can result in incorrect results from BI tools, undermining planning and analysis efforts, and exposing the organization to significant regulatory and compliance risks. It might take several business-reporting cycles to recognize and correct the errors caused by inaccurate data, which compounds the core issue. Misaligned or unclear data can also lead to inconsistent decisions, disagreements within the organization, and a general lack of confidence in the entire BI ecosystem.

Enterprises across all industries face a variety of challenges in integrating data from disparate sources and ensuring data quality. Where does the needed information exist? How can data from multiple systems be integrated for a comprehensive, accurate, and timely view into the organization’s past, present, and future? How can the organization best represent the data so that it provides pertinent insight to information consumers? How to control access to and routinely validate data to guarantee data integrity? To ensure sound data integration and quality, enterprises need to address such concerns with a comprehensive data strategy that ensures effective integration and enterprise-wide access to key data.

While enterprises have invested heavily in applications to automate data processes, much of this investment remains unproductive. Gartner estimates that through 2007, more than 25% of critical data in Fortune 1000 enterprises will continue to be flawed (inaccurate or incomplete).¹ Data quality issues include duplicate, inaccurate, ambiguous, misinterpreted, or miscalculated data. Frequently, data about core business attributes like “Customer” or “Product” are represented differently across multiple application systems or databases. Reconciling such data to achieve a single holistic and accurate view (“data alignment”) is extremely difficult in the absence of an effective data-integration policy.

Effective data quality and integration remains a serious challenge, especially for performance-driven enterprises that demand consistent visibility into activity and need to align performance with strategy. Unless data-management processes such as integration, organization, audit, and semantic consistency can be synchronized, it will continue to be very difficult for most organizations to realize a true top-level data strategy, enterprise-wide reporting and analysis, and the possibility of reliable data-integrity.

Challenges of Data Integration

Data-integration process involves integrating clean and accurate data from various data sources, transforming it, and then loading it to target data stores. Most BI projects and data warehouses end up devoting the lion share of their resources for just this process. There are various hurdles on the way to implementing an effective strategy for data integration.

A wide range of approaches to data integration are in use today, including:

- **ETL (Extract, Transform, and Load):** Data is extracted from source systems, transformed using business rules, and loaded into a target system like a data warehouse.
- **EII (Enterprise Information Integration, or ‘federated data’) — extraction and transformation of data on demand from source systems; no loading into a target system.
- **EAI (Enterprise Application Integration):** The exchange of data between systems; this is an event-driven process.
- **MDM / CDI (Master Data Management / Customer Data Integration):** When applied specifically to Customer data, MDM is known as CDM; Customer Data Integration. MDM / CDM is the strategy for providing a consistent view of core business entities such as Customers and Products.

In this paper we discuss data integration using Extract, Transform, and Load (ETL).

Organizational Considerations

In successful organizations, data is recognized as a key competitive asset and the collection, management, reliability, and application of data is given a high priority. Unfortunately, many organizations lack an enterprise-wide commitment to data integrity. In a recent survey 67% of organizations reported having a company-wide data strategy in place. However, the general level of confidence in data quality remained flat between the 2001 and 2004 surveys—with only 37% of respondents saying that they were “very confident” about the quality of their data.

One reason for the slow progress around data quality is a lack of executive commitment to data integrity as a key
business issue—in the same survey, respondents reported that data quality was discussed only in 12% of management meetings (down from 14% in 2004). The commitment to enterprise-wide data integrity must begin in the boardroom, and communicated down the chain of command as a key strategic priority for all levels of management. A common view of data results in consistency of planning and execution across the enterprise, improved collaboration based on a “common view of the truth”, a more reliable process for regulatory reporting, and the ability to use third-party data and to distribute corporate data to third parties.

![Figure 1.](image)

In addition to an organizational commitment to data integrity, data-integration solutions must be planned and deployed with a focus on the key business and technical drivers required by the organization as a whole as well as by key stakeholders and information consumers. A data-integration solution must be deployed using the same discipline for evaluating scalability, flexibility, and Total Cost of Ownership, which is applied to any technology investment decision.

**Business Drivers and Challenges**

Data-integration needs to be viewed as a decision motivated by a realization of the business impacts of ineffective data management and inaccurate or untimely reports driving bad decisions; and the reasons for the inability to produce enterprise-wide knowledge about core business objects like Customers, Customer Service, and Regulatory Compliance.

The cost involved in data integration can be high as a result of factors like data complexity and volume, existence of different versions of data across core systems, and the complexity of business rules that govern data interpretation. Apart from the resources required for designing a data-integration solution, developing and deploying the solution can be expensive and time-consuming—integration is often the most expensive part of a data warehouse or BI project.

Four major factors are responsible for the high resource costs of data integration:

- Documenting the sources of data, and mapping data from multiple source systems to a “single source, single target” in the data warehouse.
- Developing transfer methods between data sources and the data warehouse, and schedules for data transfer subject to system demands.
- Cleansing the data and developing strategies for ensuring an ongoing set of clean and reliable data in the data warehouse.
- Deploying the data-integration process in EAI, ETL, MDM, and related tools—including those for business rules, transfer schedules, and error alerts/recovery.

The need for specialized skills for planning, problem-solving and coding the many issues that arise in the course of an ETL project frequently adds to the expenses.

However, in spite of the costs involved in dealing with the complexity and criticality of data integration, ETL projects continue to be viewed as comprehensive solutions to issues with data quality and data dispersion, and as the means for achieving a common definition of key information across an enterprise.

**Technical Drivers and Challenges**

Because business data is collected from various transaction and other specialized source systems, the data can be in various formats ranging from proprietary files to XML, and even unstructured forms such as Web pages or e-mails. Thus, the complexity of organizing data, reconciling the meanings of data from different sources, and aggregating the data into a single data warehouse is aggravated by issues related to data volumes, formats, quality, and consistency of meaning. Achieving the required level of data quality and consistency with business rules is perhaps as difficult as physically consolidating the data into a single data warehouse.

Apart from attributes inherent in the scale and nature of data, the logistics of transforming the data to formats that each data consumer will value, distributing it to multiple destinations in a reliable way, ensuring scalability and performance of both source and target systems during data transfer, and meeting the load-schedule restrictions and the
demands imposed by the globalization of business; all add to the complexity of the data-integration solution.

These issues demand a planned solution that uses technology that is equipped to handle such challenges, and is implemented at every stage of the data integration process as required by the demands of that stage.

Data Integration Using Extract, Transform, and Load (ETL)

Extract, Transfer, and Load (ETL) is a combination of software and methodology to Extract data from various source systems, Transform the data based on business rules, and then Load the data to a target system—generally a data warehouse of data mart.

In today’s business world, the ability to gain holistic insight from the mass of corporate data available provides a strong competitive advantage. Information Technology departments across all industries face a sizable challenge in harnessing the data they manage to provide actionable knowledge for strategic, tactical, and operational management of the organization. Enterprises increasingly look to data-integration strategies and ETL tools to transform data into information that can be trusted by decision makers for critical business execution today—that is, aligning activity with strategy, and planning for the future.

Elevator Pitch for Effective Data Integration:

- Reduces time required for reporting and planning cycles, saves resources and enables prompt, reliable compliance reports.
- Guarantees accurate, timely, high-value information, enabling fact-based management, collaboration among business units, and alignment of activity with strategy.

ETL tools integrate data from various enterprise data sources and transfer it to the enterprise data warehouse or data marts consistently. ETL tools also help to cleanse data and govern its quality and compliance with established business rules to guarantee consistent, reliable information. ETL enables timely, accurate, and complete information in the enterprise data warehouse to support business processes such as customer analysis, planning and budgeting, and strategic planning. A reliable data-management process also inspires confidence among information consumers that business decisions are founded on fact-based reasoning.
instead of ‘gut feel’. Finally, reliable data management encourages collaboration across the enterprise, because it is based on a “single version of the truth” that is consistent and equally understood by all functional groups.

As an example, a major global manufacturing company uses its data warehouse to both analyze the performance of operations and to predict variances in scheduled deliveries. The Company’s major data sources include orders from an order entry system, customer data from the customer relationship management (CRM) system, and vendor data from the Enterprise Resource Planning (ERP) system. Apart from these core systems, the data warehouse includes external data gathered from spreadsheets and text files containing information on vendor for subcontracted deliveries, and daily operational details.

The consumers for the Company’s BI are geographically and functionally diverse across all levels of management. A primary objective of the Information Technology department is to provide a “single version of truth” for all of these information consumers. To meet this objective, the IT department needed to:

- Provide reliable historic and current data, integrated from a variety of disparate data sources into a single data warehouse
- Ensure data quality by removing duplicate data and correcting errors
- Track and certify data changes and rules to meet regulatory standards

ETL is the core technology that meets all of these needs for integration, aggregation, reconciliation, and timely delivery to stakeholders through the Company.

Stages in the ETL Process: Profile, Cleanse, and Integrate

The Data Profile Stage (Understanding Information)

Business process automation had caused a dramatic increase in the volume of data maintained by the modern enterprise. Unless there is thorough understanding of data, data-driven initiatives are like searching for the solution without knowing the problem. The first stage in the ETL process, Data Profiling, analyzes the content, structure, contexts, and hidden relationships within and between enterprise data sources. Metadata is the collection of business rules that describe and clarify the raw data stored in millions of rows and columns. Data profiling helps to examine the structure of data in the ETL process and determine whether the information matches with the rules outlined in metadata.

Data profiling provides a complete blueprint of the data-management issues faced by an enterprise. The process includes:

- Discovering multiple source systems for data
- Establishing data patterns and formats
- Performing basic integrity checks on certain sets of data, such as Valid Numeric Range analysis
- Validating and enforcing established relationships between groups of data
- Identifying duplicate or missing data (such as duplicate or missing names and addresses);
- Conforming data to business-specific rules
Data Quality Stage (Cleansing Information)

A high level of data quality is essential for successful Business Intelligence. Consistent and correct data can provide the foundation for effective and repeatable business decisions, and create a sense of trust in the entire data-management process. On the other hand, incorrect data can lead to bad decisions and a lack of faith in the entire BI ecosystem. Data used as the foundation for BI must be clean, identifiable, standardized, matched, reconciled, and free of redundancies.

Most companies have hundreds of internal and external data sources—each with different data formats, conventions, and governing business rules. Additionally, each data source can have a variety of quality issues—such as null values, missing fields, and inconsistent entries—that can lead to an overall lack of data quality across the system. The problem of data quality grows exponentially when data sources are aggregated.

Once the structures and relationships within enterprise data sources have been understood and analyzed in the Profiling stage, the scheme for implementing data quality must be planned and implemented. The Quality stage is a critical second step in the ETL process. In this phase, data quality is achieved by standardizing, correcting, and validating data (and dealing with exceptions in an automated way, or manually for extreme quality issues).

The following steps are adopted for achieving data quality in the Quality stage.

- **Standardize** source data
- **Validate, certify, and enrich common data elements** to ensure consistency in the information
- **Eliminate duplication** by matching records across or within data sources
- **Produce the ‘best’ information** from each source, for each unique entity

Data Quality Value Proposition

- Ensures reliable Business Intelligence to support fact-based decision making
- Enables data quality audits for compliance reporting
- Validates corporate information sources and creates trust in the BI ecosystem
- Expedites data migration projects

Data Integration Stage (Transforming and Integrating Information)

The success of the data integration initiative depends on discovering the true state of data in the Profiling stage and ensuring data quality in the Quality stage. The third stage in the ETL process is transforming the data based on business rules and integrating the data from multiple sources into a single target system such as a data warehouse or data mart. A proper integration strategy enables data to be extracted from any source, in any format, and delivered to any target system within or outside the enterprise, at the right time.

The objective of effective data integration is to obtain a 360-degree view of each entity that is critical to the organization’s success, such as Customers, Products, and Employees. This level of insight from information requires access to timely, reliable, enterprise-wide data from all relevant sources. The Integration stage controls the aggregation and guarantees delivery of clean, aligned enterprise data to target databases and information consumers.

The Integration stage involves:

- **Identifying and linking** entities from different sources
- **Extraction of data** from source systems
- **Iterative validation and cleansing** (steps from the Quality stage) can be performed throughout data movement subject to established business rules
- **Creation of new data** derived from raw source data (‘persistent cubes’)
- **Ensuring high-performance** of the data transfer process, with minimum impact on resources of source and target systems

An effective integration strategy lowers capital and operating expenses by striking the right balance between dedicated resources and ongoing management costs, through right-sizing of the information management infrastructure. Additionally, effective integration reduces the BI cycle time, and reduces erroneous data—both of these factors have strong impacts on BI effectiveness and the general credibility of the entire BI ecosystem. Finally, effective data management improves utilization of existing ERP, CRM, HR, and other core systems, increasing the ROI from these IT investments.
Enterprise Information Management (EIM)—a Comprehensive Approach to Data Integration

Enterprise Information Management (EIM) is a comprehensive approach to data integration, synchronizing several information management tools into a single platform of business rules and data processes. EIM is not a single tool or method, but the combination of data strategy, methodology, and technology solutions to create a unified approach to data integration.

EIM components can operate separately or as a single set of combined capabilities—the key facet of EIM is the synchronization of all data-management processes. Data-management tools are traditionally fragmented among different tools, making it difficult to implement a comprehensive data-management strategy without an EIM approach—it is difficult to integrate data processes together into a single strategy. Semantic differences are difficult to reconcile, which creates inconsistency in enterprise data, and prevents propagation of business rules between separate tools.

In addition to the benefits of effective data integration mentioned previously, the following are also drivers of EIM:

- **Rationalize (standardize) information management** to reduce costs on account of licenses, support, development, and errors from several data management tools
- **Simplify the data-management infrastructure** to reduce complexity of BI / ETL / DW maintenance; and to scale and evolve with the Business (for example, to accommodate unstructured data)
- **Better enables real-time BI and data quality assessments**

EIM includes four integrated data functions:

1. **Data Integration** (several approaches):
   - **Enterprise Application Integration (EAI)** — The connections and exchange of data between systems; typically batch-processed or event-driven
   - **Extract, Transform, and Load (ETL)** — Extract from Source Systems, Transform to business rules, and Load into target systems
   - **Enterprise Information Integration (EII)** — Also known as ‘data federation,’ EII is real-time extraction and transformation of data from source systems without loading to a target system—data is delivered directly from a source system to BI. Typically, EII is an on-demand to a BI tool.

2. **Data Quality** — Profile, clean, validate, audit, data as mentioned previously.

3. **Semantic Consistency** — Creates a common definition of key business entities such as Customers, Products, and Suppliers. Alternatively, creates a central ‘master’ set of data about these entities (MDM or CDI).

4. **Metadata Management** — Many sets of business rules (metadata) exist in BI and ETL tools; databases, data dictionaries and so on. Effective data integration synchronizes metadata in the same way it synchronizes data elements. Synchronized and centrally-managed metadata facilitates management, and enables traceability of the ‘lineage’ of data from a BI tool back to its original source.

Customer Acceptance Criteria for Data Integration and ETL Tools

With new approaches and technology for data management being continually adopted to address evolving business challenges, effective data integration has become a top priority for many enterprises. Data Integration helps to successfully manage exponentially increasing data volumes, integrate numerous data sources, and gain a holistic view of key business entities. Consequently, customers have many acceptance criteria for data-integration tools:

- **Ease of deployment and use**: Customers want tools that are easy to learn, implement, configure, design or develop, and use. They seek to reduce training and implementation cost and time.

- **Data quality capabilities**: Customers expect data-integration tools to provide accurate data quality analysis, matching, standardization, cleansing, and the capability to measure and monitor data quality.

- **Extensive data source and target support**: Customers expect integration tools to support connectivity to a range of databases, applications, and other infrastructure components, to help minimize development of system connections and adapters.
Richness of formatting and transformation capabilities: Customers expect integration tools to support a wide range of standard and custom transformation types while handling the complexity of transforming and merging data from multiple sources.

Performance and scalability: With data volumes in enterprises growing, tools need to provide the ability to process large data volumes with minimum resource requirement and impact on connected systems.

Metadata-management capabilities: Metadata is an important component of data integration. Customers need to manage metadata pertaining to discovery, auditing, lineage, impact analysis, and interoperability with other tools.

Service-Oriented Architecture: As Web Services are gaining importance as part of Service-Oriented Architectures (SOA), customers want to deploy data-integration functionality as a service and to access other data-management functions as a service.

Platform Portability: Tools must be able to support real-time data sources such as message queues, query-response, data delivery, and Changed Data Capture (CDC). Support for seamless deployment across heterogeneous platforms—distributed networks to mainframes—is also being increasingly sought by customers.

The Return on Investment (ROI) from Effective Data Integration

Enterprises have invested heavily in Enterprise Resource Planning (ERP), Supply Chain Management (SCM), and Customer Relationship Management (CRM) systems. These investments were justified by the access and insight they would provide into information regarding customer, product and operations. Effective integration enables a higher return on investment (ROI) from these applications by aggregating data and contexts from these specialized systems to enable a 360-degree view of key business elements such as Customers, Products, and Suppliers.

Additionally, successful data integration implementation helps enterprises achieve benefits horizontally across the enterprise, such as:

- More effective cross-selling and up-selling: A single, consolidated view of the customer data helps companies better understand customer preferences, interests, and demographics, which facilitates effective cross-sell and up-sell of their products and services to customers.
- Reduced marketing costs: The availability of accurate and consistent data without the duplication of customer information considerably reduces the costs associated with marketing efforts. In a 2002 survey, The Data Warehousing Institute (TDWI) estimates an additional profit of $471,000 from an average direct-mail campaign following a data quality initiative.
- Improve back-office operation: Integrated data management can lead to an improvement in back-office operations. Unified billing; accurate revenue, accounting, and contract billings; unified credit management; and reduced mailing costs are some of the benefits.
- Reduced operating costs: Consolidation of customer data into a single source can lead to significant savings in operations costs for companies. Getting the information from fragmented data is time consuming and requires resources for aggregating. In the same TDWI survey, an average call center could save $14 million following an enterprise data quality initiative.
- Improved regulatory and compliance efforts: With a single source of facts and greater assurance that data is correct and consistent, companies are more likely to be compliant with government regulations.

Summary

Enterprises gather increasingly large volumes of data from multiple core systems such as ERP, SCM, CRM, and from Business Intelligence initiatives. This data contains valuable information about customers, service levels, activity from operations, and expected future performance. This information can deliver high value to an enterprise if it is holistic, consistent, accurate, timely, trusted by information consumers, and easy to access. An effective data-integration strategy enables excellent insight and corporate agility.
Sources
1. **Gartner**: Poor Quality Data: the Sure Way to Lose Business and Attract Auditors, Friedman, March 2006
2. **TDWI Research**: Taking Data Quality to the Enterprise through Data Governance, Philip Russom, March 2006
3. **TDWI**: The ROI of Data Quality, Len Dubois, August 2002
4. ibid.

About Unisys
Unisys is a worldwide technology services and solutions company. Our consultants apply Unisys expertise in consulting, systems integration, outsourcing, infrastructure, and server technology to help our clients achieve secure business operations. We build more secure organizations by creating visibility into clients’ business operations. Leveraging Unisys 3D Visible Enterprise, we make visible the impact of their decisions—ahead of investments, opportunities, and risks.

Our services for Business Intelligence and Data Warehousing are summarized in the following table.

<table>
<thead>
<tr>
<th>Unisys BI Services</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strategic Services</strong></td>
</tr>
<tr>
<td>- 3D-VE BI Workshops (Introduction to BI Best Practices, Trends)</td>
</tr>
<tr>
<td>- BI Assessments (BI strategy and roadmaps, infrastructure evaluation)</td>
</tr>
<tr>
<td>- Outsourcing and managed services planning</td>
</tr>
<tr>
<td>- Data Management Strategy</td>
</tr>
<tr>
<td>- Proof of Concepts</td>
</tr>
<tr>
<td><strong>Infrastructure Services</strong></td>
</tr>
<tr>
<td>- BI Platform assessment and development</td>
</tr>
<tr>
<td>- Data scalability and data warehouse development</td>
</tr>
<tr>
<td>- Performance and tuning advisory, IT optimization analysis</td>
</tr>
<tr>
<td>- Application of best practices related to data quality, security, and organization</td>
</tr>
<tr>
<td><strong>Consulting Services</strong></td>
</tr>
<tr>
<td>- BI Platform Upgrades / Migration (Migrate-to-Save)</td>
</tr>
<tr>
<td>- Platform Standardization (Standardize-to-Save)</td>
</tr>
<tr>
<td>- Infrastructure Consolidation (Consolidate-to-Save)</td>
</tr>
<tr>
<td>- Project Management</td>
</tr>
<tr>
<td><strong>Managed Services</strong></td>
</tr>
<tr>
<td>- Off-site or On-site BI Platform Management</td>
</tr>
<tr>
<td>- Hosting Proprietary R&amp;D Analytics Applications</td>
</tr>
<tr>
<td><strong>High Performance Hardware</strong></td>
</tr>
<tr>
<td>- Scale-up and Scale-out Architecture Solutions</td>
</tr>
<tr>
<td>- 64-Bit high-performance Itanium servers</td>
</tr>
<tr>
<td>- Business continuity and disaster recovery solutions</td>
</tr>
</tbody>
</table>
Contacts in the Unisys BI Solutions, Practice, Sales and Alliances team:

<table>
<thead>
<tr>
<th>BI Global Solutions Management</th>
<th>Ravi Kalakota (770-368-6303) <a href="mailto:Ravi.Kalakota@Unisys.com">Ravi.Kalakota@Unisys.com</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>BI Practice Management Team</td>
<td>Kim Ramko (615) 356-9513 <a href="mailto:Kimberly.Ramko@Unisys.com">Kimberly.Ramko@Unisys.com</a></td>
</tr>
<tr>
<td>BI Sales</td>
<td>Bob Williams (610) 407-4177 <a href="mailto:Bob.Williams@Unisys.com">Bob.Williams@Unisys.com</a></td>
</tr>
<tr>
<td></td>
<td>Bill Lawlor (630) 582-2661 <a href="mailto:William.Lawlor@Unisys.com">William.Lawlor@Unisys.com</a></td>
</tr>
<tr>
<td></td>
<td>Greg Fagen (585) 487-2459 <a href="mailto:Gregory.Fagen@Unisys.com">Gregory.Fagen@Unisys.com</a></td>
</tr>
<tr>
<td></td>
<td>John Malach (585) 487-2574 <a href="mailto:John.Malach@Unisys.com">John.Malach@Unisys.com</a></td>
</tr>
<tr>
<td>BI Global Alliances</td>
<td>Shawn McPherron (801) 947-5775 <a href="mailto:Shawn.McPherron2@Unisys.com">Shawn.McPherron2@Unisys.com</a></td>
</tr>
</tbody>
</table>

For more information, contact your Unisys representative.
Or, call Unisys today at:
(800) 874-8647 (US and Canada)
00-10585-487-2430 (other countries)

Visit our website at [http://www.unisys.com](http://www.unisys.com)