

- 1. Introduction ..... 2**
  - 1.1 Who should read this report? ..... 2
  - 1.2 ETL Facts ..... 2
  - 1.3 Assumptions ..... 2
- 2. Evaluation Results of the ETL Tools by Features..... 3**
- 3. Bench Marks Ratings on the ETL Tools..... 6**
- 4. Strength/Limitations of the ETL Product..... 9**
  - 4.1. Informatica ..... 9
  - 4.2 Data Mirror..... 11
  - 4.3 Embarcadero..... 13
- 5. Features that are NOT yet implemented in these tools..... 14**
- 6. ETL Market Trends..... 18**
- Appendix A Sample Request for Proposal..... 19**
- Appendix B ETL Magic Quadrant..... 30**
- Appendix C - Cost Comparison Tool ..... 30**

## 1. Introduction

This report outlines the features of the ETL tools in the Market as well as their strength and weaknesses. The report will also investigate a range of Extract/Transform/Load tools (ETL tools) available in the market place, detailing the Advantages and Disadvantages of each tool and assessing each tools suitability in relation to a solution.

Twenty plus market-leading products have been selected and evaluated. The products have been evaluated based on survey taken by Evaltech, Inc. in January 2003 and the related web sites including forums.

This report will also give a recommendation on the ETL Tool features that are missing in the existing Tools. These features might be part of the beta version of some products.

Sample Request for proposal is also included in this report; this will help the evaluator of the ETL Tool or the vendor of the ETL Solution in understanding the industry expectations.

### 1.1 Who should read this report?

- CIO, CEO, CTO
- Project Managers
- Enterprise Architects
- Vendors
- ETL Solution Providers
- Information Technology Consulting Companies
- ETL Developers
- Data Warehouse Architects

### 1.2 ETL Facts

- ETL and data cleansing tools are estimated to cost at least one third of effort and expenses in the budget of data warehouse.
- ETL process cost 55% of the total cost of data warehouse runtime
- Data Warehouse expenses are expected to come up to 14 billion dollars worldwide, projected sales for ETL and data cleaning tools are expected to rise to only (!) 300 million dollars.

### 1.3 Assumptions

This report is not meant to teach the data warehousing concepts, it is assumed that the reader has a very good knowledge about the data warehouse and its related terminology.

## 2. Evaluation Results of the ETL Tools by Features

Features	Transformat ion Server 4.6	PowerCenter 5.0	DT/Studio Version 1.9
<b>Data Extraction</b>			
· Distinguish data differences	Y	Y	Y
· Build a query using a Graphical User Interface	Y	Y	Y
· Run a query and filter the result set	Y	Y	Y
· Sorting of results	Y	Y	Y
· Calculate results from other fields	Y	Y	Y
· Save query for reusability	N	N	N
· Extract null values, zero values, blanks, etc.	Y	Y	Y
· Display query before execution	N	Y	Y
· Display and edit extraction commands	Y	Y	Y
· Determination of net changes	Y	N	Y
· Previewing Queries	N	N	N
· Be able to auto paste results into another application	N	N	N
<b>Data Cleansing</b>			
· Removal of unwanted duplicates	N	N	N
· Remove unwanted or unnecessary characters	Y	Y	Y
· Correct field values to match field definitions	Y	Y	Y
· Ability to replace missing data	Y	Y	Y
· Test data integrity rules	Y	Y	Y
· Storage of any calculations	Y	Y	Y
<b>Data Exporting</b>			
· Move data in a single pass	Y	Y	Y
· Extract data in any format and sort order	Y	Y	Y
· Ease of exporting data	Y	Y	Y
<b>Metadata functionality</b>			
· Help features	Y	Y	Y
· Repository for queries and reports	N	Y	Y
· Microsoft repository	N	Y	N
· Integration	Y	Y	Y
· Publish to Web	N	N	N
· Open architect	N	Y	N
· Usage statistics - job statistics- real evaluation may be in the olap tools	N	Y	N

<b>Summarization/Aggregation</b>			
· Loading of aggregate tables	Y	Y	N
· Loading of summarized tables	Y	Y	N
<b>Matching</b>			
· Match records across multiple sources	Y	Y	Y
· Merge records	Y	Y	N
<b>Transformation</b>			
· Code translation	Y	Y	Y
· Re-usability	N	Y	N
· Smart code translation	Y	Y	Y
<b>Load</b>			
· Loading capabilities	Y	Y	Y
· Loading to various platforms	Y	Y	Y
· Loading many to one tables	Y	Y	Y
· Loading one to Many table	Y	Y	Y
· Automation of loading	Y	Y	Y
<b>Performance</b>			
· Simple queries should run under 15 seconds	Y	Y	Y
· Very Complex queries should run under 15 minutes	Y	Y	Y
· Tool should be easy to install	Y	Y	Y
· Scalability	Y	Y	Y
· Generation of code	N	N	Y
· Temp storage utilization	Y	N	Y
· Exception reporting/error logs	Y	Y	Y
· Scheduling	Y	Y	Y
· Pricing	50k-80k	75k-250k	35k-75k
· Open architecture	N	Y	Y
· Does the product run as a service on the server/workstation	Y	Y	Y
· Does the product require a user to be logged on for it to be run	Y	Y	Y
· Can the product be loaded on a different box than db server	Y	Y	Y
· Maintenance after the product installed	Y	Y	Y
· Retry capability	Y	Y	Y
· Back up and restore	Y	Y	Y
· Ability to connect multiple servers	N	Y	N
· System requirements middleware	N	N	N
<b>Technical Support</b>			
· Technical support must be available during normal working	Y	Y	Y

hours and if needed a support agreement can be arranged			
· Technical support must be able to answer questions while on the phone or return the call with an answer within two hours	<b>Y</b>	<b>Y</b>	<b>Y</b>
· E-Mail support and bulletin boards	<b>Y</b>	<b>Y</b>	<b>Y</b>
	<b>Y</b>		
· Technical forums and user groups on the Internet	<b>Y</b>	<b>Y</b>	<b>Y</b>
· Web page	<b>Y</b>	<b>Y</b>	<b>Y</b>
· Cost of Maintenance	12%-18%	10%-18%	9%-18%
· Upgrading costs	<b>Y</b>	<b>Y</b>	<b>Y</b>
<b>Company Stability</b>			
· The financial condition of the company	<b>Fair</b>	<b>Excellent</b>	<b>Good</b>
· What % of revenue does the company spend on R&D	<b>25-30</b>	<b>25-30</b>	<b>15-20</b>
<b>Training</b>			
· User Manuals	<b>Y</b>	<b>Y</b>	<b>Y</b>
· CBT Materials	<b>N</b>	<b>Y</b>	<b>N</b>
· Ease of use	<b>Y</b>	<b>Y</b>	<b>Y</b>

\* DTA – Data not available

### 3. Bench Marks Ratings on the ETL Tools

Features	Transformation Server 4.6	Power Center 5.0	DT/Studio Version 1.9
<b>Data Extraction</b>			
· Distinguish data differences	6	8	6
· Build a query using a Graphical User Interface	5	8	7
· Run a query and filter the result set	7	7	7
· Sorting of results	9	9	9
· Calculate results from other fields	9	9	9
· Save query for reusability	0	0	0
· Extract null values, zero values, blanks, etc.	10	10	10
· Display query before execution	0	5	3
· Display and edit extraction commands	7	4	7
· Determination of net changes	3	3	6
· Previewing Queries	0	0	0
· Be able to auto paste results into another application	0	0	0
<b>Data Cleansing</b>			
· Removal of unwanted duplicates	0	0	0
· Remove unwanted or unnecessary characters	10	10	10
· Correct field values to match field definitions	10	10	10
· Ability to replace missing data	10	10	10
· Test data integrity rules	10	10	10
· Storage of any calculations	4	7	5
<b>Data Exporting</b>			
· Move data in a single pass	10	10	10
· Extract data in any format and sort order	10	10	10
· Ease of exporting data	4	6	3
<b>Metadata functionality</b>			
· Help features	10	10	10
· Repository for queries and reports	0	10	10
· Microsoft repository	0	10	0
· Integration	6	10	8
· Publish to Web	0	0	0

· Usage statistics - job statistics- real evaluation may be in the olap tools	0	8	0
· Open architect	0	10	0
· Use Multiple Meta Data	0	8	0
<b>Summarization/Aggregation</b>			
· Loading of aggregate tables	10	10	0
· Loading of summarized tables	10	10	0
<b>Matching</b>			
· Match records across multiple sources	10	10	0
· Merge records	10	10	0
<b>Transformation</b>			
· Code translation	10	10	10
· Re-usability	0	7	0
· Smart code translation	3	6	5
<b>Load</b>			
· Loading capabilities	10	8	9
· Loading to various platforms	8	8	9
· Loading many to one tables	10	10	10
· Loading one to Many table	8	10	10
· Automation of loading	10	10	10
<b>Performance</b>			
· Simple queries should run under 15 seconds	7	8	9
· Very Complex queries should run under 15 minutes	6	8	9
· Tool should be easy to install	8	6	7
· Scalability	7	8	7
· Generation of code	6	10	7
· Temp storage utilization	6	10	6
· Exception reporting/error logs	4	6	3
· Scheduling	5	7	8
· Pricing	5	3	8
· Open architecture	0	8	3
· Does the product run as a service on the server/workstation	10	10	10
· Does the product require a user to be logged on for it to be run	10	10	10
· Can the product be loaded on a different box than db server	10	10	10
· Maintenance after the product installed	4	2	4
· Retry capability	8	9	7
· Back up and restore	8	8	5
· Ability to connect multiple servers	0	8	0

· System requirements middleware	<b>0</b>	<b>0</b>	<b>0</b>
<b>Technical Support</b>			
· Technical support must be available during normal working hours and if needed a support agreement can be arranged	<b>10</b>	<b>10</b>	<b>10</b>
· Technical support must be able to answer questions while on the phone or return the call with an answer within two hours	<b>10</b>	<b>10</b>	<b>10</b>
· E-Mail support and bulletin boards	<b>10</b>	<b>10</b>	<b>10</b>
· Technical forums and user groups on the Internet	<b>10</b>	<b>10</b>	<b>10</b>
· Web page	<b>10</b>	<b>10</b>	<b>10</b>
· Cost of Maintenance	10%-18%	10%-18%	15%-18%
· Upgrading costs	<b>5</b>	<b>2</b>	<b>6</b>
<b>Company Stability</b>			
· The financial condition of the company	<b>3</b>	<b>7</b>	<b>8</b>
· What % of revenue does the company spend on R&D	4	6	5
<b>Training</b>			
· User Manuals	<b>10</b>	<b>10</b>	<b>10</b>
· CBT Materials	<b>0</b>	<b>10</b>	<b>0</b>
· Ease of use	<b>5</b>	<b>9</b>	<b>7</b>

## 4. Strength/Limitations of the ETL Product

### 4.1. Informatica

---

Web site: [www.informatica.com](http://www.informatica.com)

Name of Current Release: PowerCenter 5.0

#### **Strengths:**

- One of the strongest features of Informatica PowerCenter is its Meta data capabilities and creating the development environment. Using PowerCenter, graphical map and code data transformations from a common set of Meta data become easy.
- XML transformations are easy to use
- PowerCenter manages and transforms data from the legacy systems into the data warehouse at the heart of hub-and-spoke architecture.
- PowerCenter manages and transforms data from the enterprise data warehouse into data marts, enabling customized perspectives of the data as required by the business users of the data marts.
- PowerCenter manages the return of data from specialized data marts back to the enterprise data warehouse or to other data marts
- Support is one of their selling tool, their online support and user forums are loaded with WebPages.

#### **Limitations:**

- Cannot Process billions of records
- Lacks the proper built-in web logs reader
- It is not a fully object oriented compliant
- Getting process flow documentation reports out of the Meta data repository is somewhat difficult.
- Lots of bugs in the transformations like lookups and Mapplets
- Cannot pass the flat file as a table in the source qualifier, need to create a separate source qualifier and use the join transformation.
- SQL validation is not 100% fool proof.
- Mapping validation is not 100% fool proof.
- Trial Version of the product not available

- One weakness is that Informatica does not allow for easy version control in the product. Because it is so easy to make changes to processes in the tool, a version control mechanism should be provided.

**Vendor Background:**

Informatica was founded in 1993 as data movement/data integration software company. Current headquarters are in Redwood City, California. Informatica sells directly to small and large businesses and value-added-resellers in the US. A network of "International Distribution Partners" supports sales efforts all over the globe. Informatica has more than 1,700 customers in use, ranging from small "Mom and Pop" local businesses to large enterprises like EDS and Arthur Andersen.

**Package Background:**

Informatica PowerCenter® 5.0 is the industry's leading enterprise data integration platform for building, deploying, and managing enterprise data warehouses. Informatica PowerMart enables users to easily transform data from disparate enterprise systems and sources into reliable information for strategic business analysis and provides the foundation for fueling Informatica packaged analytic applications, as well as custom-developed applications built using Informatica products or 3rd party business intelligence tools. Powermart and PowerCenter are both client/server applications. It's easiest to think of PowerCenter as a superset of Powermart. Powermart allows a single server to run on a single computer. With PowerCenter, one can "network" PowerMarts together w/centralized management and share metadata between these PowerMarts. Also, some high-end options are only available with PowerCenter: SAP, PeopleSoft, high-speed access to DB2/OS390 (you can access DB2/OS390 with Powermart via DB2-Connect), integration with Trillium cleansing software and recently announced e-commerce enhancements. With both PowerCenter and Powermat, you get unlimited clients.

## 4.2 Data Mirror

---

Web site: [www.datamirror.com](http://www.datamirror.com)

Name of Current Release: DataMirror Transformation Server 4.6

### Strengths:

- Transformation Server supports real-time data replication, so the data warehouse is always fresh and fully synchronized.
- Features are in place for Transformation Server to support wireless business intelligence environments. The product can serve as a bridge among mobile devices, embedded databases, and corporate databases.
- Dark Room Operations enable users to set alarms, triggers, and notifications.
- Bridges to third-party analysis tools (Hyperion, Cognos, Oracle) are provided out of the box.
- User Exits may be customized for data-cleansing functionality.

### Limitations:

- Advanced ETL features require integration with complimentary DataMirror products. For example, MQSeries is supported through Constellar Hub, load balancing is provided through the High Availability Suite, etc. These tools are not part of a Transformation Server license and are provided at extra cost.
- Impact analysis features are not yet available.
- Data extraction from COBOL formats is not supported on non-AS/400 platforms.

### Vendor Background:

DataMirror was founded in 1993. Current headquarters are in Toronto, Ontario, but the company also has offices in North America, Europe, and the Asia Pacific. Sales and distribution offices span 30 countries worldwide. DataMirror has 320 employees and more than 1,400 customers. The company is traded publicly on the Toronto Stock Exchange (under ticker symbol DMC) and on Nasdaq (DMCX). DataMirror common shares began trading on Nasdaq on January 18, 2001.

Revenues for FY00 were \$56.9M, up 35% from FY99. Net income for this period was \$4.7M, compared to a loss of \$1.4M the year before. Revenues for 1Q01 were \$13M. Strategic partnerships include IBM and JD Edwards.

### Product Background:

DataMirror.s ETL solution primarily consists of the Transformation Server, Enterprise Administrator and Constellar Hub. Transformation Server is a data integration solution that enables users to capture, transform, and flow data in real time throughout the enterprise and across the Internet. Transformation Server supports data integration and transformation among DB2 UDB, Oracle, SQL Server, Sybase, and Pointbase across Unix, Linux, Windows NT/2000, IBM OS/400, and OS/390. Flat-file replication is also supported.

Enterprise Administrator is DataMirror.s Web-enabled, Java-based GUI for Transformation Server. Enterprise Administrator provides the environment for configuring and managing data movement/transformations from a single point of administration. Metadata management features are also provided.

Additional DataMirror components include iTransmit (for wireless administration of Enterprise Administrator), iDeliver (integration software for B2B and CRM environments), Constellar Hub (for enterprise application integration), and DataMirror.s High Availability Suite (for business operations in AS/400 environments). These tools all complement Transformation Server, enabling the product to extend functionality into markets other than data replication and ETL. These tools, however, are not part of a Transformation Server license and are therefore provided at extra cost.

DataMirror offers support for multiple data schemas: pure relational, star schema, snowflake schema, and proprietary OLAP databases working with Hyperion Essbase, Oracle Express, and Cognos Impromptu/PowerPlay. Transformation Server provides native access (non-ODBC) to database interfaces for DB2 UDB, SQL Server, Oracle, Sybase, and Pointbase. The product also comes with support for .dark room operations,. which enable users to set up alarms and alerts that can be used to notify administrators via pager, e-mail, or cell phone on certain selected events.

These features are new in Version 4.6.

Transformation Server supports three types of data integration modes: real-time data mirroring, net change (showing only what has changed since the last job), and refresh (while still active). Data filtering is supported at either the row or column level. Joins can also be created at the source level, and source-derived columns may be added as well.

Enterprise Administrator provides a display of the Transformation Server Replication Network with a list of publication servers and DBMSs. Users create .subscriptions,. which are pipelined flows of data specifying hosts, servers, etc. After subscriptions are designed, tables from the catalog may be associated. The catalog contains all tables available for replication, and is filled by the user. Subscriptions are created for each flow of data, and filters are easily applied in a check-box format. Once the source is defined, users then go to the target table to associate appropriate targets. Expressions are created from a long list of column functions. Value translations are supported here as well.

Enterprise Administrator is also used to create .user exits,. which are back-end programs that can be modified to suit specific ETL requirements.

### 4.3 Embarcadero

---

Web site: [www.embarcadero.com](http://www.embarcadero.com)

Name of Current Release: DT/Studio Version 1.9

#### **Strengths:**

- Data Change Capture
- Remote Management Using a 100% Java™ Client
- Over 1,000 Data Operation Functions
- Ready-to-use Data Transformers

#### **Limitations:**

- Installation & Configuration not easy
- Document support missing
- The help files could be expanded and simplified.
- It would be difficult to run DT/Studio on documentation alone. However, the one-on-one technical support is a very effective alternative.

#### **Vendor Background:**

Embarcadero Technologies provides a suite of award-winning products that enables organizations to efficiently and accurately architect, integrate, administer and ensure critical enterprise applications and their underlying databases. It's headquarter is located at San Francisco, CA.

Embarcadero Technologies has built a large and loyal customer base that encompasses global corporations, leading financial institutions, and government agencies around the world. Customers include Hewlett-Packard, Bank of America, Sprint, Pepsi, NBC, Steelcase, and Morgan Stanley.

#### **Product Background:**

Affordable ETL solution with a full set of capabilities to help you address your data integration needs. Data housed in disparate sources must be transformed, migrated or integrated before it can be utilized as a valuable information asset. The task of integrating data becomes more difficult as your data and applications grow in number, size, complexity and distribution. Many data integration problems are left unaddressed or go undiscovered because the tools fall short. DT/Studio equips you with everything you need to address your company's data integration needs, including those like application integration and migration that extend beyond data warehousing. DT/Studio makes ETL simple and affordable so that a broad range of data integration projects are now within your reach. DT/Studio combines visual data modeling with a visual data flow designer and an extensible Java-based ETL engine to achieve new levels of usability, scalability, and flexibility.

#### **ER Data Model Designer**

Streamline the analysis phase of your data integration projects with DT/Studio's data modeling component. DT/Studio's visual modeling capabilities, like reverse engineering, help you create comprehensive blueprints of databases and data warehouses. With a clear understanding of data sources, you are better equipped to implement reliable data integration solutions in less time, with less effort. Data models created with ER/Studio or DT/Studio can be used in either application.

### Data Flow Designer

Visually diagram and develop the flow of data through your environment from simple data movements to complex column mapping and transformations. DT/Studio applies a top-down approach to visualizing data movement, starting from the macro view of data sources and targets. Point-and-click with your mouse to easily drill down on any part of the diagram in order to reveal more detail. At the detailed level, wizards guide you through the process of creating the data flow, one step at a time to reduce errors and increase productivity.

## **5. Features that are NOT yet implemented in these tools**

### **Installation & Configuration**

- Cannot register more than one server at any given time
- Cannot access more than one repository at any given time
- Unix installations are more complicated as compare to Windows installation
- Some upgrades are really complicated

### **Repository Management**

- XML primarily covers syntax, not schemas
- Current Meta Data repositories are not flexible enough
  - Easy integration of new models (Schemas, Vocabularies)
  - Much easier development of Meta data based applications
  - User defined Meta data reports
  - Ease of finding the object within the mappings
- Fully automatic approaches to meta data integration
- Not linked to the business and information processes
- Does not reflect business change
- Difficult to track business rules that cross departments and business processes
- No efficient means of sharing common, complex structures throughout the enterprise's systems
- Should have component for data storage like Operational Data Store (ODS), Development, QA, Test and Production
- The ETL should be able to construct data hierarchies and dependencies that span process flows, producing queued environments and timetables for latency processing of particular information.

### **Designer**

- Poor presentation of the relationships between tables and the files
- No indication of the primary and foreign keys
- Cannot create the user defined transformation
- Cleansing Process
- Eliminate duplicate records transformation
- Encryption/Decryption transformation
- Proper demoralization transformation
- Union, Join and Difference transformation
- Format mismatch
- Report on the unused ports or columns in the mappings

- Look-ups against tables and source files must be responsive to messaging systems and able to dynamically add, update and delete rows on demand. In this case, if multiple streams are loading a single target table, the integration must be such that the lookups across these multiple streams can be synchronized without deadlock contention.
- Aggregations must be equally shared and synchronized across multiple running streams.
- It should allow the developer of the processes to construct message flow diagrams based on conditions and execute different messages across the processes.

- Input information must be able to be dynamically reassigned to other processes based on meta-data controlled rules designed by the business.

### **Re-Usability**

- Objects should follow object oriented programming standards
- Should be able to check-in and check-out from one folder to another folder
- Should be able to edit any object

### **Server**

- Should be able to load into multiple target types like tables and the flat file at the same time.
- Real-time data capture and loading
- Ability to start multiple tasks on multiple source systems
- Get multiple feeds going in a managed and restartable manner on multiple platforms at the same time to coalesce the data into the final tables in a third-normal-form warehouse, an operational data store, or a dimensional data mart
- Produce favorable benchmarks with 1 terabyte volumes on medium size machines
- Integrate best-of-breed database features (upserts, views, portioning and indexing)
- Implement event-driven processing and RDBMS trigger-based processing
- Facilitate messaging between streams, across streams, and in and out of other processes.
- Increase fault tolerance
- Multiple streams must be able to be dynamically consumed by a single process based on business rules.
- Once initialized, continuous build and append must be available.
- ETL processes should be distributable across all registered resources so the workload is shared.
- The ETL should manage (internally) and possibly eliminate target contention or deadlock, allowing the developers of the process to focus on the complex task of getting the NRT data in and integrated.

### **Scheduler**

- Process Flow representation
- Multiple occurrences of single jobs that flow data from source to target with in-stream transformations.
- A framework that can provide access to many heterogeneous data sources with automatic checkpoint/restart capability and automatically handles placement and partitioning of the data.
- Proper dependencies of the jobs
- Initiate auto-process recovery (based on meta-data rule sets created by the designer).
- Incorporate auto loading balancing
- Use meta-data rules based metrics generation (Run-time, CPU utilization, disk-utilization, capture and recording of all metrics).
- Each process must be capable of passing a message to other running process
- Process must be able to be defined as “never-ending”, that is until they receive a message to stop.

- Queuing mechanism must be threaded and parallel so messages between the processes are not lost due to CPU time out.

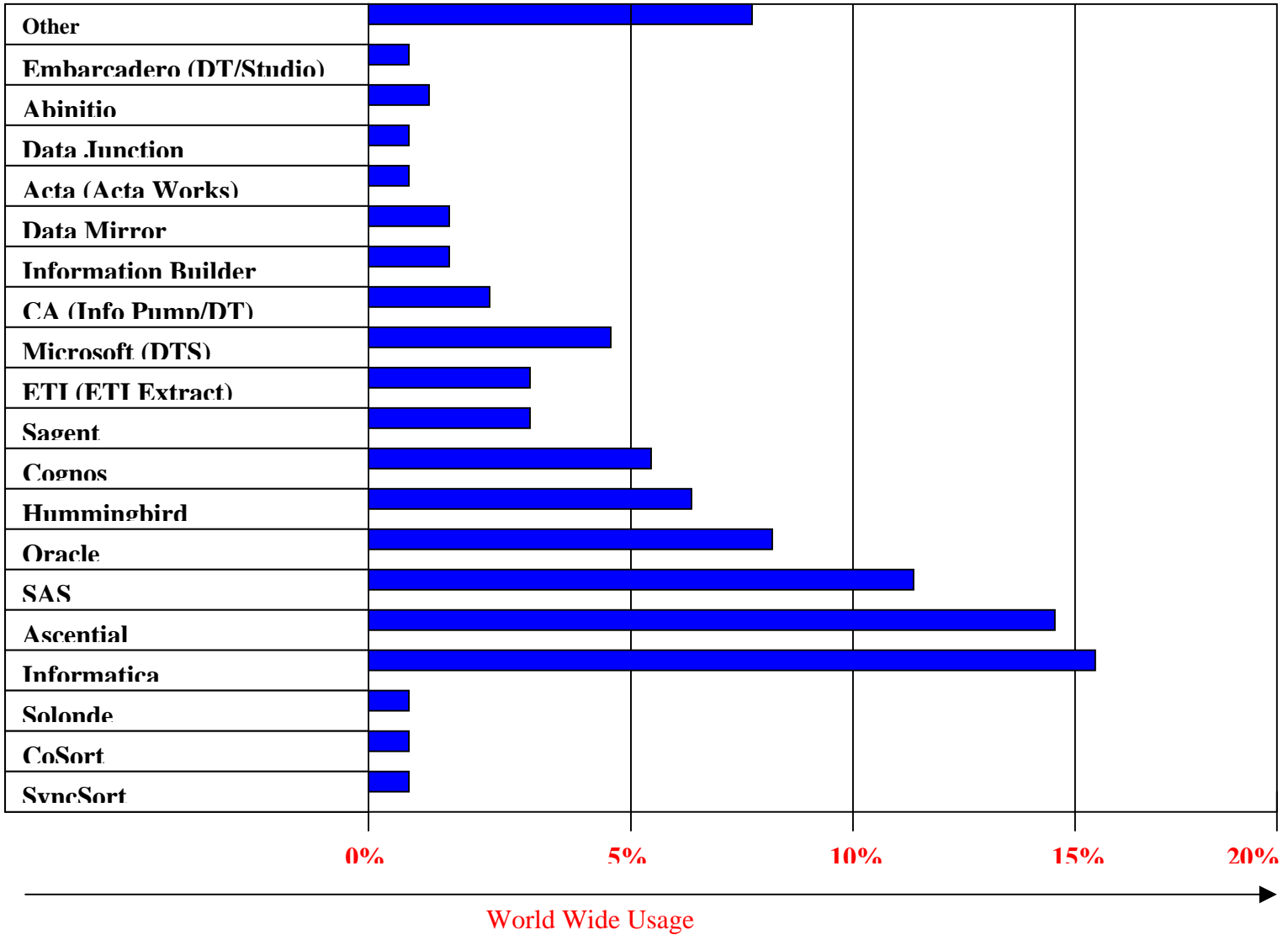
### Error Handling

- Primary Key Violation support
- Increase fault tolerance
- Checkpoints and failure recovery must be built in (or allowed to be designated) at certain points in the process. Recovery for a process must be a matter of seconds, not minutes or hours.
- Recovery should consist of returning to the most recent checkpoint.
- Wizard for email notification component
- Should create a default error handling table in the repository

### Others

- **Versioning** – Should support multiple instances of data dimensions that can be saved and used for historical reporting, comparisons, reconciliation and “what if” analysis
- **Business rule enforcement** – user-defined validations, ranging from simple data verification to organizational policies, that can be applied across dimensions in real-time or batch mode
- **Hierarchy debugging tools** – SQL-like instruments used to analyze multiple data sets
- **Property Inheritance** – a method of leveraging data, where points within a hierarchy can “inherit” information from higher points, eliminating the need to enter and store redundant data (e.g. every new customer placed within a region will inherit information associated with that region)
- **Report Writer** – builds and saves user generated reports
- **Audit Tracking** – tracks all activity by user ID and timestamp
- **Architecture** - The architecture of these next-generation ETL solutions resembles a Web portal: The location of any object is virtual, and the metadata regarding its location, characteristics, and the transformations required to present it are embedded only within the portal server, or, in our example, the active warehouse.
- **Audit** - There's no way to automate the audit and balancing procedures, which means more manual work, more overhead, and less accuracy.
- **Load utilities are unidirectional:** You can load, but you can't retrieve data. That limitation means you can't do referential manipulations (lookups from code tables) or sequential processing, such as matching records from multiple sources, on a different host. You can't join to other tables to check for data or referential integrity. And you can't look up a value in another table to determine which table a row might go to or to insert a value for a code (for example, decoding a bitmap into an indicator value or a code, such as currency type).
- **Unstructured Data** - There is a definite disconnect between unstructured source data (for example, an HTML or PDF document) and the DBMS.
- The entire ETL environment will be metadata driven.
- Incorporate configuration management capabilities
- Special wizard for Web logs for the clickstream data warehouse

### 6. ETL Market Trends



# Appendix A Sample Request for Proposal

**REQUEST FOR PROPOSAL** [Date Here]

**DATA WAREHOUSING  
Extraction Transformation and Loading (ETL) TOOL  
FOR  
XYZ, Inc.**

**XYZ, Inc. 19**

**1. Objective 19**

**2. About XYZ, Inc. 20**

**3. Response Requirements 20**

**4. Executive Summary 20**

**5. Introduction 20**

**6 Proposal Format 20**

**7 Contract Terms 20**

**8 Exceptions 21**

**9 Company Profile 21**

**10 Financial Stability 21**

**11 Billing Methods 21**

**12 Costs 21**

**13 Support 21**

**14 References 21**

**15 General Product Requirements 21**

Detailed System Requirements ..... 22

**16 Notification of Award/Contract 22**

**17 XYZ, Inc. reserves the right to 22**

**18 Liability 23**

**19 Disaster Recovery Responsibility 23**

**20 Project Schedule 23**

**General Information 23**

Phone Number..... 23

General Terms and Conditions..... 24

**21 Insurance Requirements 25**

**Appendix A 25**

## 1. Objective

This Request For Proposal (RFP) is issued for the purpose of supporting the company-wide initiative of developing data warehousing architecture to provide better access to information to foster better and more informed decision-making. The ETL tool is an integral part of the data warehousing architecture that will be used by the data warehousing professional staff to model and create target warehouse, extract data from multiple data sources, transform the data to make it accessible to business analysis, and loading multiple target data marts.

## **2. About XYZ, Inc.**

*Mission of the company*

*Executives profile*

*Awards if any*

*Number of Clients*

*Locations/Branches*

*Address*

*Employees*

## **3. Response Requirements**

Please submit your response in eight (N) copies, tabbed as shown below. (Where N is the number of copies)

## **4. Executive Summary**

Include in this tab an executive summary discussing the highlights of the proposal. Do not include pricing information in this tab.

## **5. Introduction**

Include in this tab any information your company wishes to submit about the nature of your business and primary business focus. Include information about what makes your company (not the product) different from your competitors.

## **6 Proposal Format**

**6.1 XYZ, Inc.** will be imaging all proposals. DO NOT BIND the proposal. The official name of the firm submitting the proposal must appear on the outside front cover of each binder, together with the title "RFP Data Warehousing ETL Tool Proposal".

**6.2** Each Proposal page must be numbered consecutively from the beginning of the proposal (Executive Summary) through all appended material.

## **7 Contract Terms**

Include copies of all contracts or agreements you expect **XYZ, Inc.** to sign.

## 8 Exceptions

Include in this section either a statement that you will make no exceptions to the requirements of the RFP, or a statement clearly indicating any exceptions and a statement of substitute wording for resolving the exception.

## 9 Company Profile

Please include the following information about your company; Nature of business and primary business focus, percentage of business that comes from tool referenced in the RFP, major company differentiator – what makes your company (not product) different, address for headquarters, address for office closest to **XYZ, Inc.**, contact information including - name, title, address, phone, fax, email and web site address.

## 10 Financial Stability

Each proposal must include a certified external audit statement and the 10-K report of the last corporate fiscal period for the firm submitting the proposal. Each Vendor must submit documentation indicating at least three years of experience, as of the Proposal submission date, in providing services similar to those required in this RFP.

## 11 Billing Methods

Describe in detail your billing methodology. **XYZ, Inc.** is exempt from payment of certain taxes. It is expected that invoices will not include any taxes for which an exemption applies. A tax exemption certificate can be provided upon request.

## 12 Costs

Please specify pricing for each component. Indicate if pricing would differ if only certain parts of the system were purchased by **XYZ, Inc.**

## 13 Support

Proposals must indicate the name, experience level, and length of service with the firm of the customer service representative who would be assigned to **XYZ, Inc.**. A description of the on-going support program must be provided. Describe your data mart development methodology. Describe levels and hours of support available.

## 14 References

Please list at least five (5) clients that you have done business with in the past year. Include the company's name, address, telephone number, contact name and number of years as a customer.

## 15 General Product Requirements

For the next three years **XYZ, Inc.** embarks on the company-wide initiative of developing a data warehousing architecture to provide better access to information to foster better and more informed decision-making. The ETL tool is an integral part of the data warehousing architecture that will be used by the data

warehousing professional staff to model and create a target warehouse, extract data from multiple data sources, transform the data to make it accessible to business analysts, and load multiple target data marts. [**Explain your OLTP System here**]

### **Detailed System Requirements**

**15.1** System Requirements are described in Appendix A.

**15.2** Selection Criteria A team from the [Department Names] will evaluate the proposals. Clear and concise responses are requested. After an initial review, this team may determine that further discussions and/or vendor presentations may be required. The following criteria (not listed in order of priority) will be used to select the firm:

**15.2.1** The degree to which the proposed system satisfied the **XYZ, Inc.** business requirements and management process.

**15.2.2** Overall system functionality and flexibility, including ease of use

**15.2.3** Initial cost of product and ongoing license and maintenance costs

**15.2.4** Support and training provided for implementation

**15.2.5** Corporate financial condition

**15.2.6** Ongoing maintenance and training

**15.2.7** Ability to accept XYZ's terms and conditions

**15.2.8** References

### **16 Notification of Award/Contract**

**XYZ, Inc.** will notify the successful vendor through a letter of intent. A contract will be negotiated between **XYZ, Inc.** and the successful vendor. The contract will, among other provisions, incorporate this RFP and the successful bidder's proposal. Upon execution by the Vendor and **XYZ, Inc.**, the contract will be submitted for final approval and a Purchase Order will be issued. **XYZ, Inc.** will notify unsuccessful vendors in writing.

### **17 XYZ, Inc. reserves the right to**

**17.1** Reject any and all proposals received in response to this RFP.

**17.2** Waive or modify minor irregularities in proposals received, after prior notification to the Vendor.

**17.3** Adjust or correct cost or cost figures with the concurrence of the Vendor if errors exist, and the Vendor establishes that a verifiable error occurred in the computation of the proposal.

**17.4** Adopt all or any part of a bidder's proposal in selecting the optimum configuration.

**17.5** Negotiate with selected Vendor responding to this RFP within the RFP requirements necessary to serve the best interests of the Company.

**17.6** Begin contract negotiations with another Vendor in order to serve and realize the best interests of the Company, should the Company be unsuccessful

in negotiating a contract with the selected Vendor within an acceptable time frame.

### **18 Liability**

The Company is not liable for any costs incurred by a Vendor in the preparation and production of a proposal or for any work performed prior to the issuance of a contract or delivery order.

### **19 Disaster Recovery Responsibility**

The Vendor must assume an active facilitating role in any major system failure. They must act as the primary source of technical expertise relative to the rapid re-establishment the system.

### **20 Project Schedule**

Issue RFP Date *Month, date, year*

RFP response due date *Month, date, year*

Vendor presentation *Month, date, year*

Decision date *Month, date, year*

### **General Information**

Interested firms must respond to this Request for Proposal in order to be considered.

Questions regarding this Request For Proposal must be addressed to

*Mailing Address*

*Phone Number*

*Email Address*

Interested vendors must submit an original and (N) eight copies of the proposal, in a sealed envelope clearly marked Data Warehousing ETL proposal to:

<*Mailing Address*>

Electronic submission must be sent via email

to: <*Email Address*>

### **Please Note:**

If hand carried, the office hours are 8:30 a.m. – 5:00 p.m.

Facsimile responses will **NOT** be accepted. **[Option]**

Proposal material will be treated as proprietary and become the property of **XYZ, Inc.**

**XYZ, Inc.** reserves the right to waive any irregularities in the proposals and to accept or reject any or all proposals.

**General Terms and Conditions**

**20.1** If this Request for Proposal results in an agreement to provide a data warehousing ETL tool, the following terms and conditions will apply:

**20.2** This Agreement can be modified only in writing and when signed by both parties.

**20.3** Either party may terminate this Agreement, in part or in whole, without penalty.

**20.4** Upon notification of termination, the two parties shall agree upon a turnover plan and agree upon the compensation due.

**20.5** This entire Agreement may not be assigned, sublet, or transferred without the prior written consent of **XYZ, Inc.**

**20.6** Vendor is an independent contractor, not an employee, agent or partner of **XYZ, Inc.**. Therefore, neither Vendor nor any of its employees are entitled to participate in any form of benefit or privilege that **XYZ, Inc.** extends or may offer to any of its own employees. Vendor agrees to indemnify **XYZ, Inc.** and to hold **XYZ, Inc.** harmless from and against all claims, liability, loss, damage, and expenses (including legal fees), arising from or due to any claim with respect to any part of the sales or services covered by this Agreement or any activity, Vendor, its officers, agents, or employees on or about **XYZ's** property. Vendor shall defend any such litigation brought against **XYZ, Inc.**. This clause shall survive termination of this Agreement.

During the performance of this Agreement, Vendor agrees not to discriminate against any individual because of race, color, religion, sex, or national origin, or because he or she has a physical or mental handicap or because he or she is a disabled veteran or a veteran of the Vietnam era. The aforesaid provision shall include, but not be limited to, the following: employment; upgrading; demotion; transfer; recruitment or recruitment advertising, layoff or termination; rates of pay or other forms of compensation; and selection for training, including apprenticeship, sales, and conditions of consultations regarding special needs of customers and/or clients. Furthermore, vendor will adhere to rigorously enforced principles of affirmative action regarding members of minority groups and handicapped individuals.

Services performed under this Agreement must be in accordance with all governmental laws, rules, regulations, and ordinances, including, but not limited to, OSHA, ANSI, EOA, ENCON, and Vendor certifies to this requirement. **XYZ, Inc.** reserves the right, at its sole option, to order cessation of performance in case of violation by Vendor, and **XYZ, Inc.** shall have no liability whatsoever resulting from such interruption.

Parking permits may be required for parking on campus. Rules issued by the Parking Office must be followed. If any fines are imposed on Vendor personnel, it is the responsibility of Vendor to appeal the violation and/or pay the fine.

This Agreement shall be governed by and construed in accordance with the laws of the State of **[State/Region Name]**.

## 21 Insurance Requirements

Before any services and/or work can be performed on **XYZ**'s premises, evidence of insurance in force naming **XYZ, Inc.** as an additional insured must be in the possession of the **XYZ, Inc.** Department of Risk Management, **[Address Here]**. Unless otherwise directed in writing, the following coverage's are required:

Comprehensive General Liability (including operations and completed operations) - **[\$Amount]** – occurrence, **[\$Amount]** – aggregate.

Comprehensive Automobile Liability (including owned, non-owned and hired autos) - **[\$Amount]** combined single limit.

Workers Compensation as required by law.

The vendor is responsible to maintain insurance coverage throughout the term of the agreement and/or contract. If for any reason during the term, the insurance policy is cancelled the consultant must immediately notify **XYZ, Inc.**

## Appendix A

**Product Profile:** Provide a one-page product profile for each product used in the response to this RFP. Please include the following:

- 1 Product name
- 2 Product description
- 3 Current release level
- 4 Date current release level was generally available
- 5 Projected general availability of next release level
- 6 Current product install base
- 7 Number of companies
- 8 Number of users

**Consulting:** Do you employ a refined, tried-and-tested data mart development methodology? How experienced are the individual consultants?

**Extraction, Transformation, and Loading functionality:** Provide a complete description of your proposed ETL solution. Including the key features of your ETL solution and products, as well as what makes your solution unique. Please address the following specific questions:

### Completeness

1. Provide a complete description of all phases of the data warehouse population that your proposed solution addresses: Modeling, extraction, transformation, loading, managing the data repositories and warehouse administration.
2. Describe the design/development tool set. Point and click environment versus programming.

**Development Environment**

Please address the following as it pertains to your product:

1. Ease of Use. How quickly can developers get up to speed with the tools, and how much support will they require?
2. Requires knowledge of 4GL
3. Support complete development environment, including versioning and run-time debugger
4. Ease of promoting transformation from development to production
5. Version/configuration management.
6. Easily debug transformation logic
7. Re-usable functions and automatic propagation of changes
8. Dependency analysis (assess impacts of changes). Dependency and analysis feature of source data changes
9. Specification of ETL functions using pre-packaged transformation objects, accessible via an intuitive graphical user interface
10. Ability to specify complex transformations using only built-in transformation objects. The goal is to specify transformations without writing any procedural code
11. Incremental aggregation and computation of aggregates by the ETL tool in one pass of the source data
12. Automated, slowly changing dimension support (Type I, Type II, Type III)
13. Able to join data from multiple sources. Support for concurrent processing of multiple source data streams, without writing procedural code.
14. Reuse individual transformations
15. Reroute bad records to separate target
16. Implement conditional logic for updates
17. No requirement to generate and compile source code
18. No requirement for intermediate disc files
19. Can the tool exploit outside cleansing and transformation routines and allow developers to embed their own transformation functions into it
20. Team Development. Does the toolset allow multiple developers to work on the same project concurrently, sharing ideas and results? Can the developers attach to the development environment via a LAN, WAN, or the Internet?
21. Support for data extraction, cleansing, aggregation, reorganization, transformation, calculation, and load operations, including the following functions
  - o Filter data, convert codes, perform table lookups, calculate derived values
  - o Validate data to check content and range of field values
  - o Perform procedural data cleansing functions
  - o Load cleansed data to the target data mart or central DW
  - o Re-usage of the stored procedures and functions from our production Oracle based system
  - o General transformation support (how many different transformation functions are available)
  - o Automatic generation of sequence numbers

**Management**

Please address the following as it pertains to your product:

1. Does the toolset schedule, coordinate, and execute all the steps involved in populating a data mart on a regular basis?
2. Describe data warehouse administration functions
3. Ease of installation
4. Ability to monitor and manage the runtime environment in real time
5. Support graphical job sequencer, and nesting of sessions
6. Produce audit and operational reports for each data load
7. Automatic generation of centralized metadata
8. Automatic generation of data extract programs
9. Support for the analysis of transformations that failed to be accepted by the ETL process
10. Extensive reporting of the results of an ETL session, including automatic notification of significant failures of the ETL process
11. Ability to schedule ETL sessions based on time or the occurrence of a specified event, including support for command-line scheduling using external scheduling programs
12. Ability to schedule FTP sessions based on time or event. (FTP remote flat file to Application server: UNIX or NT)
13. Access data from multiple, operational data sources: Native Interface to Oracle, Flat File interface, XML sources, Others sources
14. Restart on recovery (will the transformation process restart automatically from abnormal termination in a recovery mode, restart logic)

**Scalability**

Please address the following as it pertains to your product:

1. How scalable is the ETL tool? Can it allow scalability from 100MB data source to a 100GB data source without significantly degrading performance or requiring awkward database or platform swaps?
2. Distributed data warehouses / data marts
3. Parallel server engines (multiple application servers)
4. From one data source to multiple data sources

**Platform extensibility.**

1. Does the tool run on UNIX and NT
2. Describe options available to us in terms of moving the product from the NT environment to the Unix environment, from a product and price perspective.

**Design and Modeling:** Describe the functionality available to define the logical and the physical data models, and create indexes.

**Performance**

Please address the following as it pertains to your product:

1. Directly executable code that runs on multithreaded UNIX or NT server engine
2. One path for populating detailed transactions and aggregated data

3. Does the product require intermediate files to extract and transform relational sources or can all calculations be performed in high speed memory using multithreaded processes
4. How fast can the toolset extract, transform, and load data?
5. Supports tuning
6. In-memory data handling
7. Parallel execution of jobs

### **Meta Data**

Please address the following as it pertains to your product:

1. Meta Data Exchange architecture. Generate, manage, and maintain a central meta data repository that contains:
  - Source data definitions
  - ?Target data models
  - Transformation rules
  - ?Derived computations
2. Does the metadata repository consist of both business and technical definitions, and can it be easily browsed by end users and power users via client/server and/or Web connections?
3. Automatic generation of central metadata, including source data definitions, transformation objects, target data models, and operational statistics
4. Metadata exchange architecture that supports automatic synchronization of central metadata with local metadata for multiple end-user BI tools
5. End-user access to central metadata repository via a right-mouse click
6. Metadata exchange API compliant with COM, UML, and XML
7. Support of metadata standards, including OLE DB for OLAP

### **Security**

1. Describe the available levels of security.
2. Encrypts or hides passwords or connect strings to databases
3. Allows source and target database security to prevail
4. Requires registration of users to gain access to environment
5. Defines enforceable roles for development
6. Defines enforceable roles for operations (session execution)
7. Defines enforceable roles for administration
8. Allows for user permissions to be set by work area (or folder)

### **Your Partners**

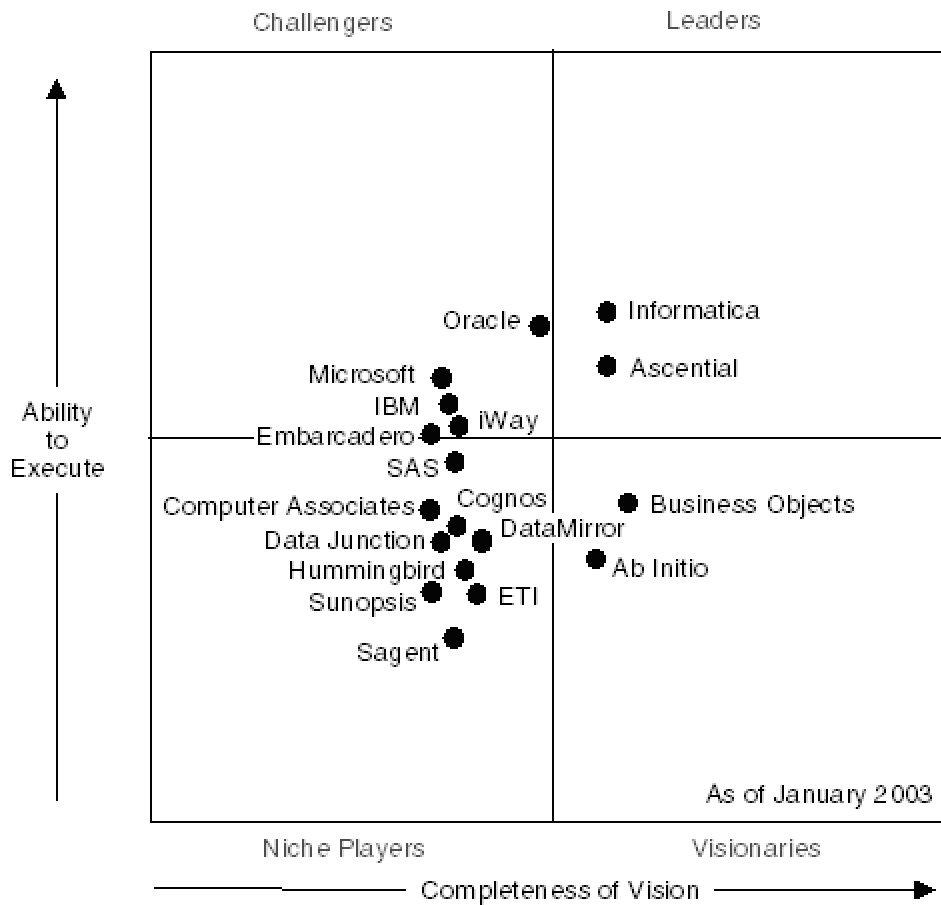
Please indicate products that are fully integrated with your solution. Please describe the connectivity and the interface requirements.

1. BI tools
2. Budgeting and Financial Applications
3. CRM applications
4. Analytic applications
5. Corporate portals

6. Others

***Price and licensing model:*** How much does the package cost, including upfront training, Per seat, support and consulting?

### Appendix B ETL Magic Quadrant



### Appendix C - Cost Comparison Tool

Please open the Excel sheet provided with this report