

SUCCESSFUL REAL-TIME BUSINESS ANALYTICS: A DATA WAREHOUSING STRATEGY

WHITE PAPER

Prepared for Informatica Corporation

September, 2002

Pieter Mimno, Independent Consultant
Mimno, Myers & Holum
www.mimno.com

Pieter R. Mimno
Pieter.Mimno@Mimno.com
P.O. Box 1095
No. Marshfield, MA 02059
Office / Fax: 781/834-3730 Cell: 781/640-3412

TABLE OF CONTENTS

1. Introduction	1
2. Challenges of the Current Environment	3
3. The New Enterprise Data Warehousing Architecture.....	5
— Source Databases	5
— Staging File	5
— Data Integration Platform.....	6
— Data Modeling Tool.....	7
— Data Warehouse	7
— Operational Data Store.....	7
— Architected Data Marts	8
4. Business-Intelligence Tools and Analytic Solutions.....	10
— Business Intelligence Tools	10
— Analytic Applications.....	10
5. Development Methodologies: Top-Down or Bottom Up?	12
6. Conclusion	15
7. About the Author	17

SUCCESSFUL REAL-TIME BUSINESS ANALYTICS: A DATA WAREHOUSING STRATEGY

Over the past decade, data warehousing has emerged as a “must-have” solution for a myriad of business challenges—from addressing the need for consistent information across the enterprise to enabling rapid response to business change. Today, companies engaging in data warehousing face a rapidly evolving array of data integration technologies, data warehouse and data mart architectures, business-intelligence tools, and analytic delivery mechanisms—all overarched by an accelerating march towards real-time functionality. The steps companies take now—and how they take them—will have a profound impact on business success going forward.

1. INTRODUCTION

The first step in building a successful data warehousing application is to identify the specific information-based problems that are causing the organization the most amount of pain. These typically include:

- The inability to extract data from multiple disparate data sources and resolve differences in data definitions
- A lack of high quality data on which to base critical business decisions
- No “single version of the truth” for business rules and data definitions
- Inability to share consistent information across business divisions
- A proliferation of non-integrated, “stovepipe” applications
- The inability to generate consolidated and reconciled financial and other business reports

Data warehousing technology is ideally suited to solving all of these problems and more. Data integration platforms can be used to access a wide range of heterogeneous data sources, resolve the inconsistencies between these sources of data, and populate target data bases. Business intelligence (BI) tools and analytic applications may be used to access the target databases to support query, reporting, and analysis of the data. Metadata, generated and maintained by the data integration platform, may be used to create a centrally managed definition of business rules and entity definitions. And it is now possible to integrate data warehousing solutions with real-time systems—including real-time clickstream analysis, real-time analytic applications, and EAI infrastructures—to drive real-time business responsiveness.

But data warehousing is not just about technologies. First and foremost, it is about meeting specific business goals—goals that need to be defined in concise, quantifiable, one-line specifications such as:

- Reduce headcount by 6% and increase workforce productivity by 20% in 12 months
- Reduce the cost of pharmaceuticals across a group of hospitals by 50% in 18 months
- Decrease the time to process small procurements from 16 hours to 4 hours via a Web-based process
- Improve productivity by 30% by automating workflow and minimizing the flow of paper
- Reduce the time spent by business analysts in resolving inconsistencies in source data from 3.5 days per week to near zero in 6 months
- Reduce the time spent by IS analysts in generating custom reports to zero in 6 months

All these example goals are eminently accomplishable with select available solutions. If a proposed new function does not directly contribute to meeting the concise goal, the implementation of that particular functionality can be deferred. But if it does line up with the goal, you should look long and hard at including it in the current development effort. Yet, above all, to ensure success, it is necessary to support even the best technology, and the implementation of the most appropriate functionality, with industry best-practices.

SYNOPSIS

This white paper examines the technology challenges that pervade today's information processing environments and details the latest data warehousing and analytic solutions—including real-time capabilities—now available to enterprises. Equally important, it also describes a proven “bottom-up” methodology for accelerating data warehouse and data mart deployment and concludes by outlining industry best-practices crucial to long-term success.

2. CHALLENGES OF THE CURRENT ENVIRONMENT

The fundamental challenge facing today's organizations is that existing hardware and software environments are not oriented toward solving the costly business problems and achieving the business goals outlined above. As shown in Figure 1, many organizations are constrained in their ability to meet business challenges due to the following environmental limitations:

- Implementation of numerous and logically inconsistent “stove-pipe” applications that support specific business functions, but cannot be integrated across the organization
- Use of resource-intensive, hand-generated procedural code to extract data from numerous sources, resolve the inconsistencies in data sources, and load target data bases
- Lack of a single, clean, consistent target database that can be used to support decision making
- Lack of a central metadata repository used to share business rules and definitions across the organization

The existence of numerous, independent, stovepipe applications makes it difficult to integrate business

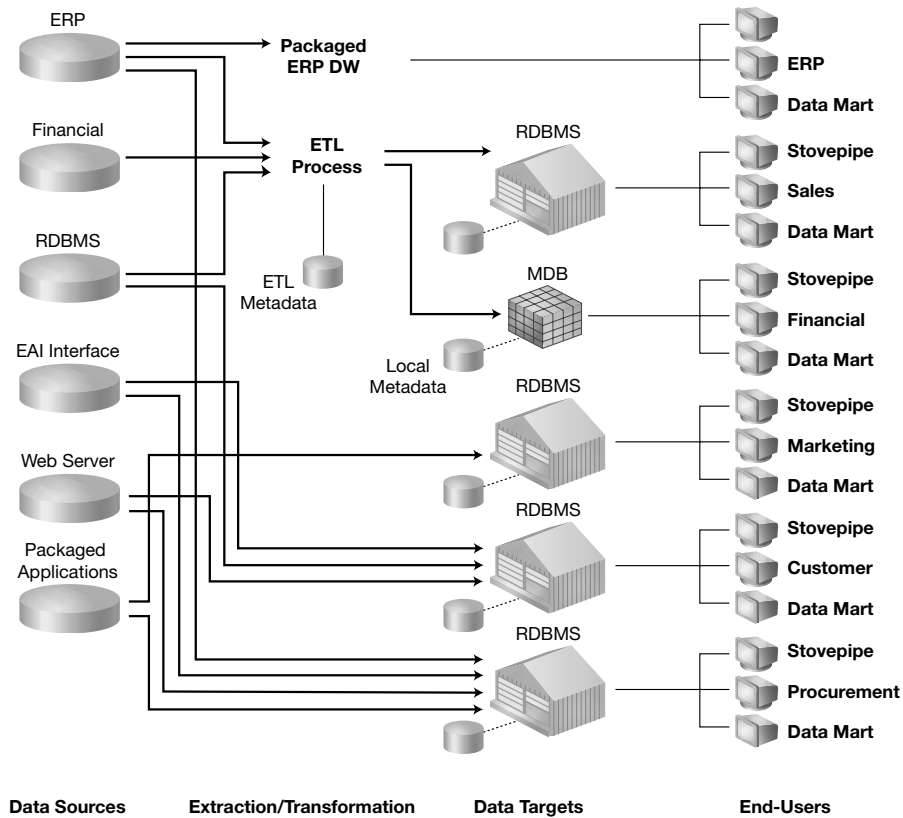
functions in response to competitive challenges. Implementation of thousands of lines of hand-coded procedural code makes it increasingly difficult to modify and enhance the extraction and transformation of source data. The lack of a clean, consistent target database inhibits the ability of business analysts to make informed decisions. Finally, the lack of integration of application modules with a central metadata repository makes it difficult to generate consolidated, reconciled reports across business units.

In the aggregate, these environmental limitations severely inhibit the ability of the organization to respond to competitive challenges. The proliferation of procedural code and the lack of metadata integration make it increasingly difficult to enhance existing applications in response to changing business requirements. And—as applications grow in size and complexity—response times to queries just get slower, reporting functions prove to be inflexible, and IS finds it more and more difficult, if not impossible, to respond rapidly to requests for enhancements in functionality.

Industry-leading data integration platform: Strengths of Informatica PowerCenter

- **Native interfaces to data sources.** PowerCenter supports native interfaces to DB2 MVS, IMS, VSAM, AS/400, relational, SAP R/3, PeopleSoft, Siebel, Web logs, XML, flat files, and real-time data sources, including IBM MQ-Series and Tibco message queues. PowerCenter also supports bulk loaders for target databases, including DB2 UDB, Informix, MS SQL Server, Sybase, Teradata MultiLoad, Essbase, SAP BW, PeopleSoft EPM, XML, IBM MQ-Series, flat files, and ODBC. In addition, the platform can be used to perform procedural data cleansing, identify anomalies in the data, transform data, compute summaries and aggregates, compute derived data, and load data at high speed into target databases.
- **Scalability to a wide range of platforms.** PowerCenter is scalable across multiple server platforms, databases, ERP and CRM applications, and eBusiness data sources. It runs on a range of platforms from Windows platforms to Unix servers. An organization can start with a modest initial configuration to support a single data mart and grow incrementally to a complex, high-performance, distributed enterprise environment.
- **Central management of enterprise environment.** PowerCenter is unique in its ability to provide centralized management of multiple, distributed data integration platform engines, synchronized with global metadata. This capability can be used to provide integrated support for multiple, geographically distributed data integration platforms, or multiple data warehouses in a Federated architecture
- **End-to-end, packaged analytic applications.** The Informatica Analytics Delivery Platform provides consistent, consolidated metrics for Customer Relationship Analytics, Web Channel analytics, Business Operations Analytics, and Supply Chain Analytics. Delivery of metrics is via a Web Dashboard display. A mobile option provides delivery of metrics via wireless and voice dashboards. Using its real-time data access capability, PowerCenter can be used to support a closed loop, real-time analytic solution
- **Intuitive design interface.** The Designer component of PowerCenter provides an intuitive, drag-and-drop interface to design source-to-target mappings and complex transformations. It incorporates an extensible library of transformation objects at a high level of significance. These high-level transformation objects enable complex data transformations to be specified in a codeless development environment
- **Metadata integration across all components of the architecture.** PowerCenter supports automatic generation of central metadata and integration at the metadata level with a family of business-intelligence tools, data modeling tools and analytic applications. A rules-based, metadata-driven engine generates and manages a central metadata repository containing definitions of source data, target data models, transformation rules, derived data computations, etc. This architecture provides a common source of definitions to integrate all components of the enterprise data warehousing architecture and prevents the development of “stovepipe” data marts that cannot be integrated across the organization.
- **High throughput using parallel transformation pipelining.** To meet the demands of shrinking update windows, PowerCenter provides high performance out-of-the box via support for multiple engines, parallel concurrent data streams, parallel transformation pipelining, extensive multi-threading, and special provisions for aggregator and join transformations.

**Figure 1: Typical Env.; No Data Integration Platform
“Stovepipe” Data Marts; No Central Metadata**



3. THE NEW ENTERPRISE DATA WAREHOUSING ARCHITECTURE

All these problems can be overcome through the use of technology, some of which was not available until recently. This new technology includes powerful, second-generation data integration platforms, new business-intelligence tools, metadata integration techniques, incremental aggregation methods, hybrid OLAP/ROLAP technology, advanced analytic applications for everyday business users, and a growing integration of data warehousing with real-time operational systems.

Figure 2 illustrates a data warehousing architecture based on the latest improvements in technology and best practices in data warehousing. Components of an enterprise data warehousing architecture include source

databases; staging files; the data integration platform; data modeling tool; the data repositories which may include a data warehouse, an operational data store, as well as architected data marts; business-intelligence tools; and analytic applications and their delivery platform.

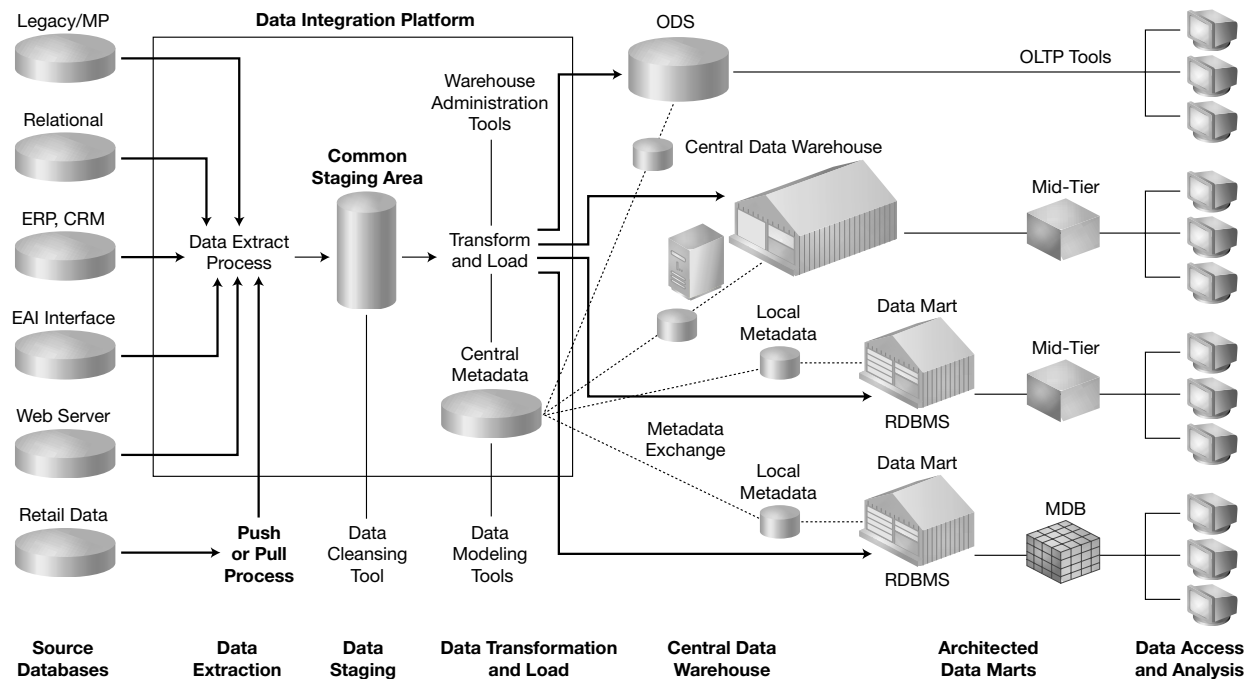
SOURCE DATABASES

Data for the data warehouse is accessed from multiple data sources, including legacy files, relational databases, ERP and CRM sources, Web log files, real-time message queues, and external information sources.

STAGING FILE

The staging file is used to accumulate data from real-time operational data sources, or stage data from multiple data sources during the nightly data extraction, transformation, and load process.

Figure 2: Enterprise Data Warehousing Architecture



DATA INTEGRATION PLATFORM

Powerful, second-generation, data integration platform tools, such as Informatica PowerCenter, are used to automatically generate executable applications, saving time and expense required to write and maintain procedural code. This is the essential platform that populates the data warehouse and associated data marts with desired information. Hence it is important that the integration platform supports a broad range of functions, including the following critical capabilities (see sidebars for more information):

- Native interfaces to data sources, including legacy, relational, ERP, CRM, XML, flat-files, and real-time transactional data (see next bullet)
- Integration of real-time data from transactional sources—including continuous real-time sessions to capture streaming data from Enterprise Application Integration (EAI) message queues and zero-latency pipelining to enable immediate processing and integration of real-time data with the historical information in the data warehouse
- Scalability to a wide range of supported platforms, including Windows and Unix environments
- Central management of distributed enterprise environment
- End-to-end, packaged analytic applications and analytic delivery platform
- Intuitive design interface for specification of data mappings and transformations
- Meta data integration across all components of the architecture
- High throughput using concurrent data streams and parallel transformation pipelining
- High speed, secure, data transmission from remote sites
- Advanced features, including versioning, incremental aggregation, debugging, reusable objects, impact analysis, slowly changing dimensions, and multi-byte, Unicode character sets
- Support of standards, including XML, COM, J2EE, and UML

Informatica Metadata Exchange

Informatica Metadata Exchange provides a rapid and easy way to import metadata from leading entity relationship data modeling tools and ERP applications into the Informatica Metadata Repository. Its use radically reduces the effort required to validate database system metadata and thus enables developers to concentrate on defining transformation and extraction processes. Specific Informatica Metadata Exchange products exist for CA Platinum Erwin, Oracle Designer, and Sybase PowerDesigner, among others.

DATA MODELING TOOL

The data modeling tool is used to specify logical and physical models for target databases. Data modeling tools that integrate with PowerCenter include Erwin from Computer Associates, PowerDesigner from Sybase, Inc., and Oracle Designer from Oracle Corporation. Physical data models specified by the data modeling tool are imported to the data integration platform repository using an integrated interface.

DATA WAREHOUSE

The central data warehouse is a database that stores large amounts of detail data (individual atomic transactions) and lightly summarized data that spans multiple subject business areas. An important requirement of the data warehouse is to capture and store a complete history of detailed, transaction-level data for multiple business areas within the organization. The data warehouse is an efficient facility to store a complete record of clean, consistent, atomic transactions accessed from all data sources.

During data update windows (which in today's environments are becoming shorter and shorter), the data integration platform is used to capture detailed transactions and changes to reference files from all databases that feed the data warehouse. The data integration platform is used to stage, cleanse, and transform the data, and to prepare it for loading into the data warehouse. The data integration platform loads the data warehouse by appending the record of detailed transactions for the current day to the record of transactions for previous days, producing a complete history of transactions. The efficiency with which the data integration platform carries out these tasks has a critical bearing on the overall success of the data warehouse project.

Queries that require access to atomic-level transaction data are directed to the data warehouse by the end-user tools. Queries directed to the data warehouse include requests for data that are not available in higher level, pre-computed aggregates, as well as queries that require access to atomic transactions across business areas, such as market basket analysis, fraud analysis, and scoring of individual customers.

OPERATIONAL DATA STORE

An Operational Data Store (ODS) is also a database construct, but it differs from the traditional data warehouse in that the ODS consolidates data from multiple source systems and provides a near real-time, integrated view of volatile, current data. An ODS may also be used as the real-time database for purchased application packages, such as financial packages, claims processing applications, etc. The architecture shown in Figure 2 supports both a data warehouse used for decision support functions and an operational data store (ODS) used for online operations. It is important to ensure the physical separation of operational data in the ODS from decision-support data in the data warehouse. Again, the efficiency of the data integration platform in handling bulk inserts and trickle-feed real-time and near real-time data inserts is paramount.

Powering the move to real time:**PowerCenterRT**

The first data integration solution to capture and process up-to-the-second transactional data, PowerCenterRT addresses the need for next-generation data integration—integration that enables real-time access to business information and metrics to drive performance management, activity-monitoring initiatives, and real-time analytics.

With PowerCenterRT, real-time data can be extracted continuously and in real time directly from EAI message queues and integrated with the historical “time-slice” information in the data warehouse. Real-time, always-on sessions integrate streaming data from EAI message queues in a continuous manner. At the same time, zero latency pipelining enables immediate processing of the real-time information. As many tasks as possible are performed in parallel—i.e.: parallel reads, writes and aggregations—and advanced caching techniques are used to reduce processing time. In addition, the transactional integrity mechanisms employed by various EAI message queues are supported.

When paired with the packaged metrics of Informatica Applications and the real-time alerting capabilities of the Informatica Analytics Delivery Platform, PowerCenterRT enables information users to receive critical information as soon as events occur in their transactional systems. Thus armed, users are able to make more timely and meaningful decisions.

ARCHITECTED DATA MARTS

Data marts contain subject-specific information supporting the requirements of end users in individual business units. Architected data marts integrate at the metadata level with central meta-data generated and maintained by the data integration platform.

Typically, data marts store summarized or aggregated data, rather than detailed transaction-level data. Aggregated data represents pre-computed answers to common queries. Data marts can provide rapid response to end-user requests if most queries are directed to pre-computed, aggregated data stored in the data mart. An important responsibility of data warehouse administrators is to identify candidates for aggregates that can be pre-computed during the nightly update cycle and stored in data marts. If a pre-computed aggregate is available in a data mart to satisfy a query, the response to the query is almost instantaneous. However, if an aggregate is not available in the data mart and the query must be directed to detail data in the data warehouse, the response to the query may be much slower.

Leading data integration platforms, such as PowerCenter, support an incremental aggregation function that is used to pre-compute aggregates as part of the data update window at night. Aggregates are computed in a single pass as data streams through the data integration platform server. At the end of the update cycle, the data integration platform loads the pre-computed aggregates into the target databases for individual data marts. Pre-computed aggregates are stored in data marts in the form of star schemas for relational databases, and in the form of indexed arrays for multidimensional databases.

For many projects, the target database for the initial development effort will consist of a single architected data mart. Following a bottom-up development methodology (described below), additional data marts can be implemented incrementally, one business area at a time. Future development phases may include implementation of a data warehouse and an ODS. In this scenario, the data integration platform loads the target database for the data mart with both transaction-level data and pre-computed, aggregated data. Later, if a decision is made to implement a central data warehouse, the data integration platform is used to move the complete history of detailed, transaction-level data from one or more data marts to the data warehouse.

Target databases for data marts may utilize either relational database management systems (RDBMS), specialized multidimensional database management systems (MDB), or a combination of both database technologies (Hybrid). Relational target databases are used to support data marts that require analysis of large amounts of data in an ad hoc, unpredictable, query environment. Multidimensional databases (MDBs) are often used to support specialized data marts that require high-speed access to moderate amounts of data (under 200 GB of summarized or calculated data) in a well-defined, predictable query environment. Financial and budgeting applications may require multidimensional databases to provide rapid response for complex, cross-dimensional calculations, as well as write-back capability for budgeting and “what-if” analysis functions. Representative multidimensional databases include Essbase from Hyperion Solutions, Inc. and Microsoft OLAP Services (a component of SQL Server 2000).

4. BUSINESS-INTELLIGENCE TOOLS AND ANALYTIC SOLUTIONS

Business intelligence tools and analytic solutions represent the interface between the user and the data warehouse. Today's business-intelligence tools are used largely by power-users and trained analysts to provide reports and forecasts to top management. Analytic applications, on the other hand, do not require specially trained users, are deployable across all levels of an enterprise, and can be applied immediately to specific day-to-day business issues. Both business intelligence tools and analytic applications draw on information that has been sourced from multiple systems and integrated by the data integration platform. The use of a business-intelligence tool or an analytic application is typically not an either-or choice. Both paradigms can provide considerable value depending on the application and can be used to complement each other's functionality.

BUSINESS-INTELLIGENCE TOOLS

Off-the-shelf, business-intelligence tools are used to specify queries, access pre-calculated reports, create ad hoc reports, and analyze data using drill-down and On-Line Analytical Processing (OLAP) functions. These tools typically access aggregated data from individual data marts and detailed, transaction-level data from the data warehouse. Data warehouse administrators generally provide end users with a choice of approved business-intelligence tools to support a wide range of end-user requirements. Regardless which business-intelligence tools are selected, it is important that they provide identical functionality in both client/server and Web environments. Compared to client/server deployment, Web deployment of business-intelligence tools provides the additional benefit of easy maintenance.

Business-intelligence tools are used with both RDBMSs and MDBs to support high-speed access and analysis of data in the target database. An important function supported by many of these tools is On-Line Analytical Processing (OLAP), which enables business analysts to access and manipulate business-oriented information from the data warehouse using multidimensional analysis

techniques. Categories of end-user query, reporting, and analysis tools include:

- General-purpose report writers and managed query tools
- Desktop OLAP tools that support managed query, reporting, and light OLAP processing by providing multidimensional views of relational data
- Relational OLAP (ROLAP) tools that enable power users and users of analytic applications to access large amounts of relational data and support complex OLAP processing
- Multidimensional OLAP (MOLAP) tools for financial analysts that provide multidimensional views of multidimensional data arrays and support for high-speed OLAP processing
- Data mining tools used to uncover hidden patterns in large data sets and support for predictive modeling.
- Data visualization tools used to display complex data in graphical form for human pattern recognition analysis

Multiple categories of tools may be required to support the full range of data warehousing requirements of an organization. However, for the initial data mart, the recommendation is to utilize a simple, low cost, off-the-shelf OLAP tool to support the query, reporting, drill-down, and OLAP processing needs of business analysts and general business end users. Such an OLAP tool can provide direct access to summarized and aggregated data stored in the target database for the data mart, as well as drill-through access to detailed, transaction-level data stored in the data warehouse.

ANALYTIC APPLICATIONS

Relatively new on the scene, analytic applications are a somewhat different animal from business-intelligence tools. While enterprises will continue to rely on business-intelligence tools, many are finding that complementing these capabilities with analytic applications increases the overall value of enterprise information and drives sizeable productivity improvements across functional areas.

Informatica Applications and Analytics Delivery Platform

Informatica Applications is a suite of analytic applications focused on the immediacy of information, its broad deployment, and its direct applicability. Among available solutions, only Informatica Applications analytic software can provide integrated real-time business views across an entire enterprise, from front office to back office, from customers to suppliers.

Informatica Applications consists currently of five separate but integrated applications:

- *Informatica Customer Relationship Analytics*
- *Informatica Web Channel Analytics*
- *Informatica Business Operations Analytics*
- *Informatica Strategic Sourcing Analytics*
- *Informatica Supply Chain Analytics*

Each module is designed to provide the type of real-time insights that executives and managers need to guide timely tactical and strategic decisions. Users define the specific information they need according to their business function and interact with the information (analyze trends, engage in what-if scenarios, etc.) through personalized analytic dashboards and scorecards. Dashboards can include active indicators (e.g.:

“the average time it takes for call center representatives to handle calls”), graphs and report links. Collaborative capabilities can be built in, including sharing of personalized reports, metrics and files. And automated drill-down paths and workflows can be implemented to quickly discover root causes and facilitate rapid insight.

Going further, Informatica Applications works in concert with the Informatica Analytics Delivery Platform to give users intuitive access to critical data regardless of where each user resides or works. The Informatica Analytics Delivery Platform provides real-time point-of-work delivery of analytic information, dashboard interfaces, and personalized alerts via Web, wireless (WAP, PDA, text pager, RIM pager), and voice recognition technologies. For example, an alert can be defined that notify a supply chain manager that rejects have reached a pre-defined threshold for a particular vendor. The alert can go out to the manager in real time, via phone, email, or pager so that immediate informed action can be taken.

In short, Informatica Applications and the Informatica Analytics Delivery Platform put decision support capabilities into the hands of everyday business decision-makers to enable the type of everyday-decisions that drive a company’s performance and define its success.

Analytic applications deliver key metrics and precisely tailored information directly to casual users and decision makers in a business context, without requiring any specialized analytic expertise. Instead of waiting for reports to be sent to them by trained analysts and power users, analytics-enabled managers themselves use business problem-specific, web-based dashboards and scorecards to evaluate key performance metrics on a continual basis. Analytics applications can also provide analytic workflows that guide managers quickly and consistently through their business decisions.

In addition, being relatively new, analytic applications are built from the ground up to support Web delivery. Users personalize when, where and how they want their information delivered and dashboards and metrics can be delivered to a wide array of information devices, including PCs, PDAs, WAP enabled devices, and text pagers, as well as voice recognition interfaces. Personalized time and threshold-based alerts can be specified and delivered in real time to a broad array of devices.

5. DEVELOPMENT METHODOLOGIES: TOP-DOWN OR BOTTOM-UP?

Given the number and complexity of the components that comprise the new enterprise data warehouse environment, there should be no surprise that there is more than one approach to building a data warehouse. The approach that is taken will have a profound impact on the velocity and cost of the overall project, both in its initial implementation and going forward.

Let's say that you want to build an enterprise data warehouse that will ultimately support 10 to 12 data marts, a central data warehouse, and perhaps an ODS. The first issue is whether to use a top-down or bottom-up development methodology. The traditional top-down approach typically requires a substantial long-term effort to interview potential users of the data marts, document user requirements, and prepare a detailed enterprise data model for the data warehouse. This often involves in-depth business discovery across multiple business units, reconciliation of numerous differences in entity definitions and business rules, and months of work to specify an enterprise data model for the data warehouse. Hence the top-down approach requires a lengthy, expensive, up-front development effort that synthesizes the requirements of multiple business units in order to define a model for the central data warehouse. In the current business climate, the top-down approach is likely to fail because it requires a large, up-front development expense and defers ROI.

BOTTOM-UP DEVELOPMENT PROCESS

A more successful strategy is to use a bottom-up methodology that builds the data warehouse incrementally, one business area at a time. The bottom-up development methodology may be used to build a data mart for a specified business area, such as sales, marketing, finance, etc., within a 90-day timebox. The bottom-up approach uses Rapid Application Development (RAD) techniques, rather than top-down Information Engineering (IE) techniques. Although the development effort is focused on building a single data mart, the data mart is embedded within a long-term enterprise data warehousing architecture that is specified in an early phase of the development methodology.

In bottom-up development, the assumption is that the long-term data warehousing architecture will be implemented incrementally, one business unit at a time. The development of more complex components of the architecture, such as a central data warehouse and an ODS, are deferred until later stages of the project. The incremental development effort is kept under control through the use of logical data modeling techniques (E-R diagrams that incrementally expand to an enterprise model), and integration of all components of the architecture with central metadata, generated and maintained by the data integration platform.

Components of the architecture include multiple data sources (legacy files, RDBMSs, flat files, spreadsheets, ERP, CRM, Web server log files, real-time message queues, etc.), a data integration platform, a central metadata repository, a metadata exchange architecture, a data modeling tool, target databases, and BI tools. The initial development effort is a 90-day project resulting in the delivery of a fully functional data mart for a specified business area. The 90-day development effort begins on the day that the data integration platform, target database, and BI tool are installed successfully.

The bottom-up methodology, which is derived from well-proven Rapid Application Development techniques, typically incorporates the following 12 steps:

1. Identify business drivers, sponsorship, risks, and ROI. Conduct a survey of high-level business managers to identify specific, painful business problems, e.g. the organization is losing its competitive advantage, customers are moving to competitors, management has little insight and control over costs, consistent information cannot be shared across business units, etc.
2. Survey user needs and identify desired functionality. Unlike the top-down approach, the survey of user needs is short (one day per business unit or less) and leads to the specification of a top-level data model for each business unit that will utilize a data mart. The top-level data models are then synthesized to identify common data sources, facts, dimensions, transformations, etc.
3. Design the long-term, enterprise data warehousing architecture. Following the survey of user needs, a workshop is convened whose function is to define the long-term vision for the data warehousing application, i.e., what will the architecture of the data warehouse look like 2 to 3 years in the future?
4. Define functional requirements for the initial subject area. A second workshop is convened to bring together the business users and development team for the first data mart to be implemented. Deliverables of the workshop include a preliminary project plan for the development effort, functional requirements for the first data mart, budget, required skill sets, etc.
5. Research and select data warehousing components and tools. Following research, a short list of potential tools is defined, including data integration platforms, data modeling tools, and BI tools. Selection of the finalists in each category may require a Proof of Concept test. Following selection and installation of the tools, the 90-day timebox is entered (steps 6-9 below)
6. Design target data base. A data modeling tool is used to model the target data base for the initial data mart. Modeling of the target database proceeds through three steps: design of an entity-relationship diagram, then a logical dimensional model, and finally a physical model of the database schema for the target data base. The physical database schema is imported into the metadata repository of the data integration platform

7. Build data mapping, extraction, transformation, and data cleansing rules. The data mapping and transformation rules are defined first in natural language, and then implemented using only the transformation objects supplied with the data integration platform. The objective is to avoid using procedural languages to code any of the extraction, cleansing, transformation, or load processes
8. Build aggregation, summarization, partition, and distribution functions. The data integration platform is used to compute aggregates in one pass of the source data, using incremental aggregation techniques
9. Complete development of the initial architected data mart, using an exact subset of the enterprise data warehousing architecture. The deliverable at the end of the 90-day timebox is a fully functional data mart for the initial business area
10. Build additional architected data marts. Additional data marts are built by a primary development team using common templates and components, such as conformed dimensions, common transformation objects, data models, central metadata definitions, etc.
11. Expand to an enterprise architecture, including a central data warehouse and an optional Operational Data Store. Development of the central data warehouse and ODS are deferred until they are clearly required. A central data warehouse is often required when detailed, atomic data from multiple data marts must be accessed to generate cross-business reports and analyses
12. Maintain and administer data warehouse. A secondary team may be used to enhance and maintain completed data marts. The primary team transfers transformation templates, data models, conformed dimensions, metadata, etc. to the secondary team to simplify the enhancement and administration of completed data marts

The bottom-up approach has the advantage that it requires little up-front investment and builds the application incrementally, proving the success of each step before going on to the next step. The first deliverable of the bottom-up approach is a fully functional data mart for a specific business area. Subsequent data marts are delivered every 90 days or less. Complex components, such as the central data warehouse or the ODS, are deferred until they are clearly required, the team has accumulated sufficient experience, and the project has demonstrated significant ROI. In the bottom-up approach, the central data warehouse and the ODS are not on the critical path and may be deferred to a later development phase. The use of logical data modeling and meta-data integration techniques ensure that all components of the application remain integrated, without the requirement for a central data warehouse.

The bottom-up methodology has been used for numerous successful applications. However, implementation of the methodology depends on several critical success factors, including a dedicated implementation team, consulting help at the beginning of the project, backing of a business manager who is hungry for a solution to a painful business problem, E-R data modeling, and integration of all components of the architecture with central metadata.

6. CONCLUSION: BEST PRACTICES IN DATA WAREHOUSING

Two things about data warehousing: One, data warehousing environments are not becoming simpler. They are complex undertakings and are likely to remain so. Two, data warehouse environments are becoming steadily more strategic to the success of the enterprise. They are not only being used for business intelligence but are also starting to take a key operational role as real-time integration technologies and real-time analytics are integrated into the mix to drive real-time decisions and actions. Consequently, it is more important than ever to implement proven best practices so as to avoid delays, excessive costs, and business disappointments as a project goes forward. Based on long experience, the following best practices are recommended to build data warehousing applications:

- Ensure that the data warehouse is business-driven, not technology-driven
- Define the long-term vision for the data warehouse in the form of an enterprise data warehousing architecture
- Avoid “stovepipe” data marts that do not integrate at the metadata level with a central meta-data repository, generated and maintained by a data integration platform
- Do not build “virtual” data warehouses that access data directly from source environments and have no target database
- Buy, don’t build data warehousing components
- Create a hub-and-spoke architecture using a data integration platform to access data sources and populate the central data warehouse, data marts, operational data store, and analytic applications
- Do not code extraction, transformation, and load functions by hand using COBOL, C, C++, PL/SQL, Perl scripts, Cold Fusion, etc.
- Use a 2nd-generation data integration platform to automate the extraction, transformation, and load functions. The data integration platform should incorporate native interfaces to source and target databases, including legacy files, relational databases, as well as ERP, CRM, Web log, and real-time data sources
- Integrate all components of the data warehousing architecture with central metadata
 - Buy only components that integrate with central metadata
 - Use the data integration platform to generate and maintain central metadata
 - Generate extensible, XML-compliant metadata and LDAP-compliant directories
 - Derive 100% of local metadata from central metadata
 - Integrate multiple, networked, data integration platform engines with common global metadata. Beware of “multiple versions of the truth”
- Do not let an ERP package, such as SAP R/3 or PeopleSoft, dominate the data warehouse. Treat the ERP application as one of many data sources and maintain the ability to build custom data warehousing applications
- Use the real-time features of the data integration platform to access real-time data and support a clickstream data warehouse. Move toward implementation of real-time, closed-loop, analytic applications

- Do not load dirty source data into the data warehouse
 - Use a data cleansing tool, in combination with a data integration platform, to clean the source data and resolve logical inconsistencies
 - Look for data cleansing tools that integrate closely with the data integration platform
 - Use the data cleansing tool to resolve inconsistencies in the Customer and Employee dimensions
- Build the data warehouse bottom-up, not top-down
 - Bottom-up development maximizes ROI and minimizes risk
 - Develop the data warehouse incrementally, one business area at a time
 - Focus initially on the development of multiple data marts. Later, if necessary, develop a central data warehouse. In the bottom-up development methodology, the central data warehouse and ODS are optional; they are not on the critical path
 - Prove each step before moving on to the next step
 - Deliver a major increment of functionality every 90 days or less
- Keep the incremental, bottom-up development effort under control through use of logical data modeling techniques and integration of all components of the architecture with central metadata, generated and maintained by the data integration platform. Create logical and physical models of the target databases with a data-modeling tool. Do not assume that OLTP data modeling experience applies to data warehousing
- Anticipate scalability and performance issues. Provide high performance through the use of pre-computed aggregates:
 - Use an automated tool to identify candidates for aggregates
 - Use the incremental aggregation feature of a data integration platform to pre-calculate aggregates in one pass of the source data
 - Use the aggregate navigation functions of a BI tool or a database to re-direct queries to pre-computed aggregates
- Use hybrid tools, multidimensional query languages, and a persistent multidimensional cache to improve the performance of the business intelligence tool and target database. Move to hybrid OLAP tools as soon as they are available
- Deliver information from the data warehouse over the Web
- Use a business-intelligence oriented corporate portal as a gateway to both structured and unstructured information
- Use analytic applications to “go places” where business-intelligence tools cannot—for example, to enable casual users to clearly see what is happening now in a business, in an operational sense
- Make sure to meld analytic application implementations with management functions so that analytics become an integral part of how users do their job
- Combine the use of business-intelligence tools and analytic applications to maximize the value of enterprise information
- Do not code reports by hand
 - Use an automated Business-Intelligence tool to support query, reporting, OLAP analysis, drill-down, and drill-through functions
 - Support ad hoc reporting, “push” reporting (reports are pushed to users on occurrence of user-defined events), and “pull” reporting (named, template-driven reports are pulled to the user)

7. ABOUT THE AUTHOR

PIETER R. MIMNO

Mr. Mimno is an independent consultant and has been actively working as a computer software professional for over 40 years. Areas of expertise include data warehousing, eBusiness intelligence applications, corporate portals, and development methodologies. He has consulted extensively on subjects related to computers and information technology and has had direct technical responsibility for the design, implementation and management of large-scale projects in fields as diverse as space systems, data warehousing, and financial management systems.

In the area of data warehousing, Mr. Mimno specializes in the selection of data warehousing components and support for all phases of data warehouse development. He assists clients to identify the business goals for the data warehouse, define user needs, specify the data required to meet those needs, and determine the size and scope of the data warehousing system.

Mr. Mimno provides consulting services in support of all phases of the planning and implementation of data warehouses, including training sessions in data warehousing technology, user interviews, definition of functional requirements, and support for the evaluation and selection of appropriate components and tools, including data integration platform tools and BI tools. Mr. Mimno also supports all phases of the development effort,

including specification of data models, source-to-target mappings, transformation rules, loading of target databases, periodic project reviews, and management of the development process.

Educational qualifications of Mr. Mimno include a B.S. in Physics and an M.S. in Aeronautics and Astronautics from the Massachusetts Institute of Technology. He has worked with computers throughout his career as a programmer, manager, and independent consultant. For the past ten years, Mr. Mimno has worked as an independent consultant. Recent consulting efforts have included specification of corporate data warehouse architectures incorporating multiple architected data marts, support for the implementation of data warehousing applications, selection of high-productivity data warehousing tools, and utilization of best practices to build data warehousing applications.

In May, 2001, Mr Mimno merged his international consulting practice with Mark Myer's and Knute Holum's implementation practice to form Mimno, Myers & Holum. The firm has established a reputation for delivering cost-effective data marts within a 90-day "time box" using a bottom-up development methodology.

More information on Mr. Mimno's data warehousing experience and on Mimno, Myers & Holum can be found on his Web site at www.mimno.com

ABOUT INFORMATICA

Informatica Corporation is the leading provider of business analytics software that helps Global 2000 organizations monitor, manage, and optimize the performance of key business operations across the enterprise. Today, more than 1500 customers worldwide have chosen Informatica to power their analytic solutions—resulting in millions of dollars in cost savings and productivity gains.

For more information, call 1.650.385.5000, or 1.800.970.1179 in the U.S., or visit the Informatica Web site at www.informatica.com

Pieter Mimno, Independent Consultant
Mimno, Myers & Holum
P.O. Box 1095
No. Marshfield, MA 02059

Office / Fax: 781/834-3730
Cell: 781/640-341
Pieter.Mimno@Mimno.com
www.mimno.com