

Data Integration in a Service-Oriented Architecture

A Strategic Foundation for Maximizing the Value of Enterprise Data

WHITE PAPER



This document contains Confidential, Proprietary and Trade Secret Information ("Confidential Information") of Informatica Corporation and may not be copied, distributed, duplicated, or otherwise reproduced in any manner without the prior written consent of Informatica.

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Informatica does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice.

The incorporation of the product attributes discussed in these materials into any release or upgrade of any Informatica software product—as well as the timing of any such release or upgrade—is at the sole discretion of Informatica.

Protected by one or more of the following U.S. Patents: 6,032,158;5,794,246;6,014,670;6,339,775;6,044,374;6,208,990;6,208,990;6,850,947;6,895,471; or by the following pending U.S. Patents: 09/644,280;10/966,046;10/727,700.

This edition published November 2005

Table of Contents

Executive Summary
Examining the IT and Business Drivers Behind SOA
Is SOA Real? Progress to Date and Remaining Challenges
Common Data Dilemmas in Service-Oriented Architectures Today
Taking SOA to the Next Level: Service-Oriented Data Integration
SOA: An Ideal Data Integration Ecosystem
Anatomy of Service-Oriented Data Integration8
Universal Data Access
Metadata Repository and Services
Ensuring Enterprise-Class Deployment
The Practical Effect: Two Business Examples
Customer Data Integration
Order Entry and Fulfillment
The Migration Towards SOA
Focus on Incremental, Business-Driven Projects
Consider Informatica Enterprise Data Integration in a Service-Oriented Architecture

Executive Summary

As business and IT managers assess strategies for service-oriented architecture (SOA), it is crucial to recognize the role that data integration plays in enabling an SOA to deliver on its potential. Many early SOA initiatives have focused on high-level application integration that abstracts business logic to more effectively broker processes, messages, and services, and enable cost-effective reusability.

At the granular data level, however, many SOA initiatives leave unresolved the issue of heterogeneous data that varies by format, semantics, and hierarchies among multiple applications. The result can be cost and time overruns to reverse-engineer data integration mechanisms, limitations in the functionality of the SOA, and inconsistent data across the enterprise.

Consider a hypothetical example. Company A had acquired Company Z. To understand the new revenue and customer dynamics of the merged company, the CIO and IT architects designed an SOA-based management reporting system meant to show the top 100 customers across the two entities. That data was stored in disparate legacy applications.

Once the system went live, though, the information was virtually useless. It was full of contradictions, redundancies, and missing data. How did this happen? The reason: the SOA lacked a data integration foundation. With an enterprise data integration platform in the mix, IT could have cleansed the data, resolved conflicting relationships and hierarchies among customer entries, and presented a uniform, accurate view of the top 100 customers from both Company A and Company Z.

This white paper examines how an enterprise data integration platform enriches a service-oriented architecture, and how an SOA provides an ideal framework for implementing data integration technology across the enterprise. After reading this white paper, you'll understand:

- . The IT and business drivers behind SOA
- The progress made to date and the remaining challenges of SOA
- · How to take SOA to the next level with service-oriented data integration
- · The anatomy of service-oriented data integration
- Approaches to migrating toward SOA

The convergence of data integration and service-oriented architecture can reduce IT complexity, ensure data consistency, and drive business agility.

Examining the IT and Business Drivers Behind SOA

In theory, SOAs provide great efficiency and economies of scale. In practice, however, SOAs have swallowed millions of dollars of investment in IT infrastructure while failing to satisfy business demands for reduced costs, greater agility, and competitive advantage in increasingly fast-paced global marketplaces.

SOAs themselves are not the problem. The fundamental problem is the complexity of IT infrastructure and the absence of an integrated architectural foundation that enables interoperability among silos of heterogeneous applications built up over years and even decades. In many organizations, well over 50 percent of the IT budget is devoted to building and maintaining points of integration among legacy systems specific to finance, supply chain, customer management, and other mission-critical functions. ClOs and IT managers are confronted with:

- High costs for IT development, deployment, and maintenance
- · Poor IT infrastructure flexibility in meeting new business demands
- · Inconsistent, inaccurate business data across the enterprise

Integrating such fragmented architectures has emerged as one of the greatest challenges facing IT and business, with bottom-line business implications. Business prosperity increasingly depends on a global view of customers, suppliers, products, and partners—an ideal not achievable without integration.

Facing brutal competition and brisk merger and acquisition activity, businesses in many industries struggle with data fragmentation. In retail and manufacturing, multiple order entry or order fulfillment systems and poorly integrated supply chains are common. Financial services and insurance companies struggle to deliver services consistently over multiple channels, including the Web, call centers and retail offices. Virtually every business is challenged to capitalize on cross-sell and up-sell opportunities, and meet the data integrity requirements of Sarbanes-Oxley and other regulations.

WHAT IS AN SOA?

A service-oriented architecture is a design approach to expose business-relevant application and data services to a variety of users and business processes in a flexible and extensible manner across the enterprise.

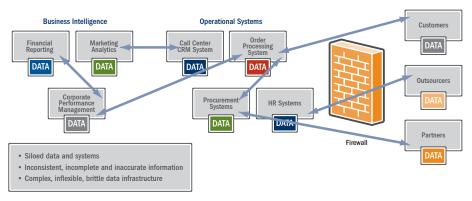


Figure 1: Data Is Fragmented Across the Enterprise, Resulting in a Complex, Brittle Data Infrastructure

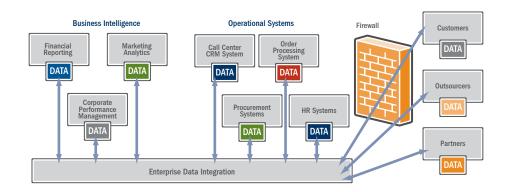


Figure 2: The Solution to Data Fragmentation Is an Enterprise Data Integration Infrastructure

Service-oriented architecture offers an elegant remedy. Though not a new concept, SOA has gained widespread acceptance with the advent of open, industry-standard protocols such as Extensible Markup Language (XML), Simple Object Access Protocol (SOAP), and Web Services Description Language (WSDL).

These standards offer a highly flexible layer of abstraction that can reduce development time and cost by liberating developers from writing low-level middleware "plumbing" and promoting reuse of components. They provide greater ease of use and have seen greater uptake than earlier architectural approaches, such as object-oriented DCOM and CORBA. At its most basic, a Web services-based SOA will cover:

"Loosely coupled" services: A layer of abstraction between the technical service implementation and client (e.g., a Web portal) that eliminates the need for customized, "tightly coupled" interoperability

Leverage of IT assets: Component-based services may be wrapped and reused for deployment across multiple projects and applications to reduce development time and cost

Use of open standards: Web services standards such as XML, SOAP, and WSDL provide the interoperability that enables an SOA to work across heterogeneous platforms

Is SOA Real? Progress to Date and Remaining Challenges

There has been a lot of hype about SOA. Heated debates continue about whether it's new and whether it's real. Real progress has been made in areas in which there was early focus, such as composite application development via service componentization and business process orchestration. Benefits have included faster development of business application functionality and integrated business processes that span multiple organizations and systems.

Many early SOA projects have been implemented using enterprise application integration (EAI) technologies and J2EE- and .NET-based middleware, including integration broker suites, message-oriented middleware, enterprise service buses, and application servers. These application integration-centric integration technologies abstract the business logic in enterprise applications and legacy systems and enable it to be accessed as discrete application services.

Common Data Dilemmas in Service-Oriented Architectures Today

However, EAI and other application integration-centric technologies are not designed to address data integration issues. SOA implementations that lack a data integration foundation have often resulted in massive hand-coding efforts to redress granular data complexities underlying the business logic, resulting in further proliferation of brittle point-to-point connectivity. Others may be handicapped in data functionality, or exacerbate problems of inconsistent data across the enterprise.

To avoid costly setbacks and implement an SOA that provides integration at both the application level and the data level, IT decision makers should define business requirements for data integration in an SOA, scrutinize the capabilities of SOA solutions to deal with complex data issues, and recognize functional distinctions between EAI, data integration, and other SOA technologies.

In particular, IT architects should consider five key elements:

Data Semantics: The business context behind data definitions for concepts such as customer address, product category, or employee type

Data Quality: Improving the accuracy and consistency of the "dirty data" common in disparate applications and legacy systems

Data Governance: Data and metadata lineage, management, documentation, reporting, and auditing tools that help satisfy Sarbanes-Oxley and other regulatory requirements

Data Access: Broad reach into structured, semi-structured, and unstructured data in hierarchical and relational databases, mainframe systems, files and documents, and applications

Bulk Data Processing: Support for processing large data volumes, including changed data capture (updating only changed data for faster performance)

A service-oriented architecture offers a framework for EAI and enterprise data integration technologies to complement each other, with EAI orchestrating processes and transactions, and a data integration platform executing complex data integration functions. Without data integration technology as a foundation, an SOA suffers limitations in its ability to fully access and leverage data, and deliver on its potential as a transformative architecture for IT and business.

Trends in Enterprise SOA Implementation

In many enterprises, SOA deployments are growing beyond early adoption to a more pervasive presence, according to a survey conducted in 2005 by IT analyst firm AMR Research of 134 IT professionals at companies of more than 1,000 employees.

In a 2005 AMR Research report titled "Service-Oriented Architecture: Survey Findings on Deployment and Plans for the Future," AMR found that 21 percent of organizations have deployed SOA today, with another 53 percent planning to implement or evaluate SOA components within the next 24 months. Early adopters start SOA projects focused first on internal business processes. Forty-four percent of current SOA deployments are focused on internal business processes. Data integration accounts for 76 percent of the initiatives. It also revealed that among survey respondents:

- 60% have partial or no integration
- 68% of integration is hand-coded
- 50% have no or limited data standardization

Leading companies that have deployed SOA recognize data integration as an important foundational element of SOA. Companies must build a solid data services layer, otherwise SOA benefits will never be realized.

Taking SOA to the Next Level: Service-Oriented Data Integration

From its roots as a single-purpose ETL (extraction, transformation, and loading) tool for data warehousing, enterprise data integration technology has evolved over the past 10 years to become a comprehensive platform to access, transform, and integrate data from a wide variety of heterogeneous systems, as well as integrate data in real time through interoperability with Java Message Service (JMS) and other messaging systems. The technology has a proven track record of dealing with extremely complex data integration issues and huge volumes of data in order to provide a consistent, accurate view of information to the business.

Today, data integration technology has evolved to supply the rich data-level functionality required for an optimal SOA and adheres to the three key principles of SOA:

- Loose coupling: Data integration services in an SOA allow a client application to access business-relevant data wherever it resides, in whatever form, in a consistent and accurate manner across the enterprise
- Leverage of IT assets: Component-based data integration services may be wrapped and reused across multiple systems without extensive rework
- Use of open standards: Support for XML, SOAP, and WSDL enables data integration technology to interoperate seamlessly throughout the SOA stack as both a provider and consumer of data integration services

In doing so, the data integration engine handles under the covers the complex data-level "heavy lifting" that maximizes the business relevance and utility of the information:

- · Reconciles disparate semantic definitions
- Executes complex transformations
- · Ensures data quality and consistency across systems

SOA: An Ideal Data Integration Ecosystem

A service-oriented architecture offers an ideal framework for implementing a common data integration approach across the enterprise. In an SOA, data integration technology takes advantage of the layer of abstraction that enables its components and services to be wrapped and reused without extensive hand-coding.

Migrating to an SOA environment also presents the opportunity to readily connect disparate instances of data integration technology with Web services protocols, JDBC/ODBC, and JMS.

For instance, it's not uncommon for organizations to have data integration software deployed

on multiple servers for separate purposes, such as data warehousing, synchronization among applications, or batch transfer of mainframe flat files.

With separation of business logic and underlying data, the open architecture of the SOA encourages leveraging and reusing these data assets by delivering them as componentized data services.

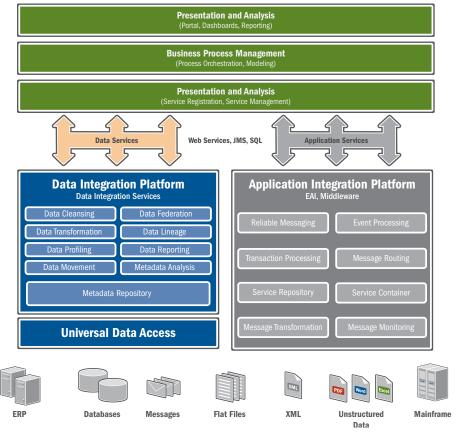


Figure 3: With a Service-Oriented Enterprise Data Integration Architecture, All Back-end IT Assets Are Accessed through Data and Application Services over Standard Web Services Protocols

SERVICE-ORIENTED DATA INTEGRATION

With universal data access, a metadata repository and services, and data integration services, data integration platform furnishes the robust data integration services and functionality essential to an SOA.

Anatomy of Service-Oriented Data Integration

A sound data foundation for an SOA may be found in an enterprise data integration platform that offers mature and specialized technologies engineered expressly to address data issues. This platform is comprised of three functional components: a universal data access layer, a metadata repository and services, and a Data Integration Services Engine. This trio operates in concert to coordinate and deliver a range of data services.

Universal Data Access

The ideal data integration platform would provide pre-built connectivity to a wide variety of packaged applications, mainframe systems, relational databases, and semi-structured and unstructured data. It should provide nearly unlimited data access via traditional physical as well as virtual data integration approaches, while minimizing the cost and complexity of accessing data regardless of where it resides.

The maturity of today's leading data integration platforms helps to greatly accelerate the design, deployment, testing, and maintenance of data access to sources including:

- · Packaged applications (SAP, Oracle, J.D. Edwards, Siebel, PeopleSoft)
- Mainframe/midrange systems (IMS, VSAM, DataCom)
- Relational databases (Oracle, Microsoft SQL Server, IBM DB2)
- Unstructured data (Excel, PDF, Lotus Notes, PowerPoint)
- Semi-structured and industry-specific data (XML, EDI, ACORD, HL7, FIX, Swift)

Universal data access should cover not just sources of data, but also flexible methods of access that suit business objectives:

- · Scheduled (traditional access for bulk movement of data)
- Real-time (near-instantaneous access on an event-driven or request/response basis)
- · Changed data (only data updated since last access)
- On-demand (federated access to data when it is requested by a client or user)

Metadata Repository and Services

The data integration platform should extend beyond data to its metadata—the "glue" that describes data values and their semantic meaning. It should provide a drag-and-drop user interface that enables developers to rapidly build business logic, processes, and transformations for data and make them available for reuse. It should have at its core a scalable metadata repository that stores and manages data models, transformations, workflows, and other artifacts to provide:

- A mechanism to reconcile data semantics among disparate systems
- · Reduced development time and cost
- Data lineage for reporting, auditing, and data governance

The metadata repository serves as a universal data interaction framework that brokers the translations between high-level service definitions and more granular data definitions and mappings. A key value is abstracting the logical model from the actual implementation of the data integration logic, which allows for better management of the business rules and transformations that form the data integration solution.

A data integration platform is unique in providing rich metadata management and analysis vital to understanding the genesis and lineage of enterprise information, determining the ripple effects of changed data, and pinpointing weak links in the data infrastructure. An SOA offers organizations that have neglected metadata an opportunity to reconcile the business meaning of data, to better understand how data and processes are derived and the interdependencies between them, and to exploit metadata for greater accuracy and efficiencies.

Ensuring Enterprise-Class Deployment

At the heart of a data integration platform is a high-performance engine for delivering data integration services, which offers a variety of flexible data delivery mechanisms and scalability for large-volume data transformations and movement over multiple concurrent sessions. Leading data integration platforms will provide:

- · Conventional batch mode movement
- Changed data capture (moves only updated data for improved performance)
- · Real-time data capture and movement
- · Built-in data partitioning for multi-CPU parallel processing
- · Support for distributed, multi-node grid systems
- · High availability, fault tolerance, and failover

The data integration engine also fortifies security at the data layer through robust authentication and authorization, support for identity management protocols such as LDAP (Lightweight Directory Access Protocol) and ActiveDirectory, and RSA data encryption.

The Dynamic Delivery of Data Integration Services

With universal data access, metadata repository, and data integration services, the data integration platform furnishes the specialized data services and functionality essential to service-oriented data integration:

Data Profiling: Thorough, accurate information about the content, quality, and structure of data in virtually any system

Data Cleansing: Resolving missing data fields, correcting conflicting data, and managing data relationship and hierarchy resolution to ensure the quality of data supplied to the business

Data Transformation: Creating a consistent view of data irrespective of its source, structure, and business semantics

Data Federation: Federated access and integration of data when data cannot be copied, or to avoid data latency or cost issues of building a physical data store

Data Movement: Flexible, "right-time" delivery of data wherever it is needed

Data Lineage, Data Reporting, and Metadata Analysis: Reporting, analysis, and auditing of data and metadata for improved visibility and control

The Practical Effect: Two Business Examples

Two compelling areas in which a data-enabled SOA can deliver tangible benefits to both business and IT are customer data integration, or CDI, and order entry and fulfillment.

Customer Data Integration

In mergers and acquisitions, companies are pressed to quickly leverage a new customer base to establish credibility, build loyalty, and capitalize on cross-sell and up-sell opportunities. Let's say two newly merged divisions of the same company, each with its own customer base, are looking for cross-sell opportunities between the divisions. The company would be faced with the daunting task of consolidating vast stores of customer names, address, business history, and other information stored in incompatible and geographically dispersed systems.

A Web services-based SOA offers an expedient means of integrating the systems between these two divisions. However, unless a data integration platform is at the foundation, the merged entity faces the risk of inconsistent and "dirty" data—conflicting customer names, addresses, and business history. A data integration platform supplies the data access, cleansing, transformation, and movement services required for an optimal transition and functions as the nucleus of its CDI initiative.

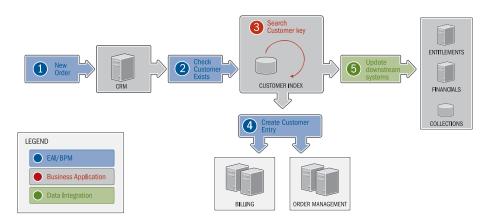


Figure 4: Data Quality, Data Movement, and Data Transformation Services Are Executed for SOA-Based Customer Data Integration Solutions

Order Entry and Fulfillment

According to a recent industry survey, the average company today has multiple order entry systems and as many order fulfillment applications, the majority of them poorly integrated or hand-coded with limited reusability. For example, a manufacturer envisions an SOA environment to better orchestrate ordering and fulfillment, improve developer productivity, and offer customers a more efficient ordering experience.

In this scenario, the manufacturer recognizes that its SOA should extend beyond process orchestration to better integration of the data underlying the various operational systems in the order entry, supply chain, and order fulfillment processes. It implements a data integration platform to better synchronize the order and customer data across these various systems in an optimized manner to minimize ordering bottlenecks, satisfy top revenue-generating customers, and streamline the overall manufacturing, distribution, and accounts receivable functions.

REAL-WORLD FINANCIAL SERVICES CDI

A large financial services company is using Informatica in an SOA that supports a customer data integration (CDI) solution. The customer has several ERP systems along with a CRM system that contain customer information.

To ensure a single, consistent view of customers, the company implemented an SOA that uses an EAI bus and the Informatica data integration platform. Access to the various data elements and application logic in the systems is achieved through discrete services. In this architecture, the EAI system drives business processes and checks if a customer exists in a customer data repository. Informatica populates back-end ERP applications with new or updated customer information in the data format for each individual system.

Updating the back-end system involves reconciliation of the customer data across the various systems, and complex transformations to ensure that the data is delivered in the appropriate formats, and with the correct relationships and hierarchies.

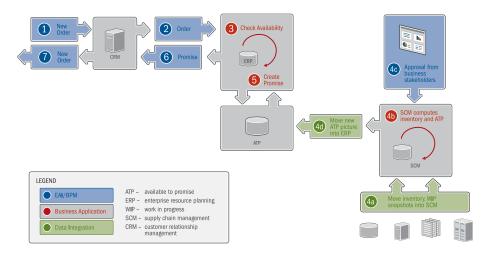


Figure 5: Demand Data Movement and Set-Based Transformation Services Are Executed for Order Management Solutions in an SOA

The Migration Towards SOA

A service-oriented architecture is not built in a day. It is, after all, an architectural transformation—and as such compels a strategic assessment of IT infrastructure components, an alignment of IT resources and business objectives, and education among IT professionals. Organizations are also likely to confront organizational, cultural, and ownership issues that transcend departmental boundaries in both IT and business.

To help ensure a smooth transition to a SOA, IT may benefit from focusing on incremental projects with high ROI and by establishing an integration competency center (ICC).

Focus on Incremental, Business-Driven Projects

Beginning the road to SOA is all about picking the right first projects. It's important to identify something that is focused in scope, relatively fast to implement, delivers a business benefit, and offers reuse of assets developed in follow-on projects. As follow-on projects reuse the same assets and leverage a common infrastructure, it's important to track metrics and be able to show the reuse and cost savings to IT management, and the time savings to business executives.

Business managers should play a key role in prioritizing areas in which an SOA implementation can deliver the most rapid return on investment. IT and business should understand the existing data issues and architecture, examine the data integration objectives of an SOA, and quantify the expected ROI as well as less tangible benefits, such as enhanced data quality.

A phased approach to the data-enabled SOA offers an opportunity to develop best practices, generate lessons learned, and improve collaboration among business and IT.

Establish Integration Competency Centers

There has been a major shift towards coordinating data integration projects across the entire enterprise. This shift has resulted in the emergence of integration competency centers (ICCs). ICCs are an organizational approach designed to increase agility and lower costs by creating a central pool of skilled resources, promoting reuse, sharing best practices, and establishing common processes and standards for integration. They are particularly relevant in the evolution towards SOA and shared services.

To support ICCs, a data integration platform must support universal data access to address:

- The full breadth of data integration initiatives across an enterprise
- Mission-critical performance, scalability, and availability
- Global IT team collaboration, capturing data workflows and logic as metadata for reuse

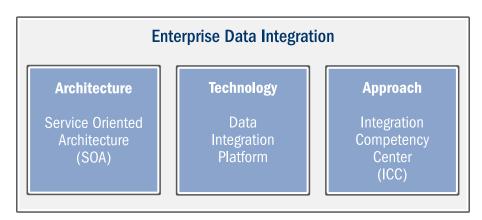


Figure 6: An ICC Is the Proven Approach to Building and Managing SOAs in an Enterprise

Informatica supports and encourages organizations to establish ICCs. An ICC serves as a center of excellence that helps IT and business managers develop reusable practices, accelerate integration projects, share resources, refine best practices, and substantially reduce costs. While the specific structure of an ICC will vary depending on an organization's needs and existing structure, ICCs have four basic models:

- 1. Best Practices: This ICC model defines processes for data integration initiatives and recommends appropriate technology, but does not own technology or engage in actual implementation activities.
- 2. Technology Standards: This ICC model standardizes development processes on common technology standards, enabling greater reuse of work from project to project. Although neither technology nor people are shared, standardization creates synergies among disparate project teams.
- 3. Shared Services: In this ICC model, processes are defined, architecture is standardized, and a centralized team maintains shared work and environment, but most development work occurs in the distributed lines of business.
- 4. Central Services: In this ICC model, standards and processes are defined, technology is shared, and a centralized team is responsible for all development work on integration initiatives.

Consider Informatica Enterprise Data Integration in a Service-Oriented Architecture

Informatica's enterprise data integration platform sets the standard for data integration across the enterprise. It is uniquely designed for enterprise-wide deployment and distributed architectures and provides an ideal foundation for service-oriented data integration by delivering:

- · Universal data access to virtually any enterprise information
- Comprehensive data integration services including data profiling, transformation, movement and federation
- Improved data quality, lineage, auditing, and security
- · Robust metadata management capabilities
- · Reusability for rapid deployment, greater IT productivity, and high ROI
- · Open, platform-neutral architecture and standards-based interoperability

Data integration is a critical component of an enterprise IT organization's SOA. To achieve SOA's goals of loose coupling and reusability, it is critical for client applications to have the ability to access business-relevant data wherever it resides, and in whatever form is required, and in a consistent and accurate manner. Data integration services, powered by a data integration platform that provides universal data access and a metadata-driven architecture, provide that capability.

Data Consumers

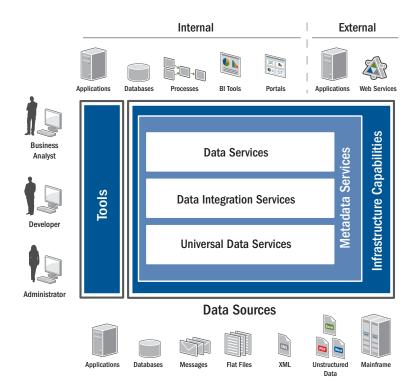


Figure 7: Informatica Platform Architecture

Companies need an enterprise data integration platform that delivers the foundation for service-oriented data integration and more importantly, enables an IT organization to evolve incrementally into the desired architecture. The Informatica enterprise data integration offering is unique because it is a single, unified platform that:

- Offers the broadest access to all enterprise data. Informatica's platform expands the definition
 of universal data access. Our platform provides virtually unlimited data access, dramatically
 reducing the cost and complexity of accessing data regardless of where in the enterprise it
 resides. No other product provides access to all enterprise data types—structured, unstructured,
 and semi-structured. No other product combines traditional physical and virtual data integration
 approaches in a single platform. Creating virtual data views without having to physically move
 data is particularly valuable when addressing the vast stores of enterprise data residing on
 mainframes and packaged applications (e.g., SAP).
- Is the only platform suitable for mission-critical, enterprise-wide data integration deployments. Informatica offers unique capabilities that empower customers to address exponential growth in their data volumes. High availability, and seamless recovery and failover capabilities in our platform eliminate single points of failure and minimize service interruption in the event of hardware and/or software outage. Our platform sets the enterprise standard for performance by adaptively applying grid resources for parallel processing as well as intelligently optimizing these resources by "pushing down" processing to a database. And no other product is faster. Our change-data capture technology allows users to not only exploit the raw power of our data integration engine, but also judiciously apply CPU cycles only towards data that has been changed.
- Provides for unparalleled productivity for global IT teams. Only Informatica offers the depth and range of capabilities for boosting productivity of not only individual developers, but for entire global IT teams. Using Informatica's unique metadata repository infrastructure, companies can easily standardize and reuse definitions across the enterprise. More companies have relied on Informatica to achieve mature implementations of ICCs than with any other technology. The Informatica platform sets the standard for enterprise-level productivity with team-based development to support distributed teams, mapping templates with integration with common modeling tools, and the ability to leverage the existing skill sets of Java developers.

White Paper

