INFORMATICA

WHITE PAPER

The Next-Generation of Data Integration: Transforming Data Chaos into Breakthrough Results

Expanding the Scope of Data Integration to Meet Emerging IT and Business Demands

This edition published February 2013

Table of Contents

Executive Summary

Data integration has traditionally meant compromise. Data architects and developers could rapidly execute a project to meet a specific deadline, but speed would come at the expense of quality. Alternatively, they could focus on delivering a quality project, but it would drag on for months longer than anticipated. Finally, teams could ensure quality and rapid delivery, but costs would spiral out of control. Regardless of which path you chose, the end result would be less than desirable. It no longer has to be this way.

Next-generation data integration changes the equation by eliminating traditional tradeoffs and enabling data management teams to execute better, cheaper, and faster projects. Focusing on people, processes, and technologies, the next-generation approach to data integration takes advantage of best practices and tools that have steadily matured with greater capabilities and reach to drive an evolution toward the ideal of an agile, data-driven enterprise.

This white paper examines the IT and business pressures that are making next-generation data integration a strategic imperative and outlines key characteristics of a next-generation data integration environment. It also provides practical guidance for making the transition and highlights the success of Informatica customers in using next-generation data integration to improve the cost efficiency, quality, and speed of integration for the purpose of decisively addressing critical business needs.

Why Is Data Integration Still So Difficult?

Despite years of building and refining data warehouses and data infrastructures, IT teams continue to struggle with high costs, delays, and suboptimal results of traditional data integration. Complexity is a key culprit. The ceaseless introduction of new data sources and data types—at high volumes and breathtaking speed—has meant that data management professionals are perpetually taking one step forward and two back.

Rising business demands are another factor. The business clamors for accurate, timely information, and yet the responsibility to deliver that data is in many cases almost entirely shouldered by an already overworked IT department. Subpar collaboration across the business and disparate data management teams contribute to delays and undermine results. The statistics are sobering:

- **Takes too long:** A survey by The Data Warehousing Institute (TDWI) found it took nine weeks to add a new source to a data warehouse in 2012—two weeks more than the seven weeks the same task took in 2008.[1]
- **Costs too much:** The analyst firm Gartner estimates data integration costs will double without adding a focus on data quality.[2]
- **Lack of trust:** A Ventana Research study found that only three in ten organizations view the data used in their analyses as always accurate.[3]
- **Poor scalability:** Gartner has calculated that 85 percent of data warehouses built in 2010 will by 2014 lack the scalability needed to support growing data volumes and complexity.[4]
- **Data is outdated:** TDWI has found that 93 percent of companies are willing to invest in real-time information for existing warehouses.[5]

[1] The Data Warehousing Institute, "2012 BI Benchmark Report" and "2008 BI Benchmark Report."

[2] Gartner Research, "Data Integration and Data Quality: Disciplines Merging, Markets Converging," February 2011.

[3] Ventana Research, "Delivering Trusted Information for Big Data and Data Warehousing: A Foundation for More Effective Decision-Making," July 2012.

[4] Gartner BI Summit, "Spinning BI Gold from Data in the Cloud," January 2010.

[5] The Data Warehousing Institute, "Next Generation Data Integration," April 2011.

The 5 Whys: Root Causes of Data Integration Inefficiency

The interrogative technique known as the "5 Whys," used in lean manufacturing, lean data integration, and other disciplines, is useful for pinpointing the root causes of data integration inefficiency. By understanding root causes, data management professionals enhance their ability to chart a course for improved results.

- **Data integration takes too long** because of the growing complexity inherent in first-generation data architectures, which lack reusability of data integration patterns and best practices.
- **Integration costs soar** as data management professionals labor with inflexible tools and architectures, while the absence of sound metadata complicates change management processes.
- **Data is untrustworthy** because of the lack of holistic data quality and governance and the inability of data environments to readily adapt to changes in data types and sources.
- **Scalability suffers** in traditional data environments as data volumes, variety, and velocity grow as the big data era unfolds.
- **Data is outdated and stale** because of the lack of real-time technologies to more frequently update data, which causes long delays in creating reports

These problems and challenges are all related, driven by the reality that data has become more fragmented while data integration has grown more complex, costly, and inflexible. Continued reliance on hand-coding—whether across the enterprise or in isolated pockets—exacerbates the problem by inhibiting reusability and draining valuable IT resources.

At the same time, pressure is mounting for IT teams to expand the scope of data integration beyond its roots in data warehousing to encompass more techniques and business areas, many related to big data. Operational data integration, master data management (MDM), business-to-business (B2B) data exchange, data virtualization, and capture of social media and other web data are some of the key focus areas. The introduction of Hadoop and other big data technologies, as well as growth in cloud-based data systems, poses an additional challenge in managing hybrid data environments.

Dramatic changes in data volume, variety, and velocity make the traditional approach to data integration inadequate and require you to evolve to next-generation techniques in order to unlock the potential of data.

Next-Generation Data Integration Defined

Next-generation data integration includes a host of technologies to allow reuse of work across the breadth of data integration projects, not just data warehousing. This includes traditional extract, transform, load (ETL), real-time data integration and replication, data virtualization, big data (Hadoop), data quality, MDM, B2B integration, information lifecycle management (ILM), messaging, and complex event processing. It is based on an agile and modular architecture that allows IT to proactively respond to changing business demands.

Next-generation data integration technologies are designed to support an extended team that includes data, integration, and enterprise architects, as well as data analysts and data stewards, while better aligning with business users. The benefit of team-oriented collaboration and tooling is an order of magnitude improvement in agility, cost, and overall quality, which enables IT to transform data chaos into breakthrough results by:

- **Streamlining development** to increase productivity and agility while lowering the cost of data integration
- **Assuring confidence** by allowing you to implement reliable processes that convert data into high quality, current information for effective decisions
- **Enabling Innovation** without barriers through easy mechanisms to harness technology advances that unleash the potential of your data

Maximizing Returns with Next-Generation Approaches

Leading Informatica customers are strategically and aggressively evolving toward next-generation data integration. By leveraging the latest Informatica technologies and aligning people, processes, and architectures, enterprises are achieving the following remarkable results:

- Doubling productivity through unified data integration that promotes reusable work and greater collaboration
- A 5X improvement in agility, acceleration of data integration projects, and better alignment with business sponsors and users
- A twofold reduction in costs from greater overall efficiency and optimized resource utilization across a variety of projects
- Twice the scalability with the flexibility to utilize multiple data processes (e.g., changed data capture, caching, and streaming data)
- Quantifiable business impact in such areas as revenue, lower business costs, customer retention, and time to market

Next-generation data integration teams are distinguished by treating their work as a “data business” within the overall enterprise, with a clear focus on serving the needs of their customers—the business. To succeed, they strive to offer products and processes that are more valuable and actionable than any alternative. A centralized Integration Competency Center (ICC) sometimes serves as headquarters for the “data business,” with strong management and oversight in promoting reusability, setting priorities, optimizing resource allocation, and measuring success.

Successful teams also employ “lean thinking” as a data integration framework, endeavoring to eliminate waste and inefficiency. Leaders recognize that the evolution toward next-generation data integration is not an overnight change but rather focuses on continuous improvement. Finally, leaders drive their efforts based on business ROI rather than merely the IT effort required.

The Next-Generation Data Integration Architecture Characteristics

A next-generation modular architecture can't be engineered overnight but instead matures as a data integration team pursues high-value, low-risk initiatives and incorporates continuous enhancements into the environment. The result is that you don't have to rip and replace your old integration architecture but can replace and upgrade it over time. Figure 1 represents a next-generation data integration architecture where:

- The outer boxes create complexity because they change frequently.
- Each radial box represents a class of systems that must be integrated, such as COTS, B2B semi-structured standards, Cloud apps, unstructured data, and mainframe computers.
- Lines to the center play two roles; the outer portion is source-dependent logic, while the inner portion is independent of the source.
- The inner cylinder changes least frequently and is based on canonical/enterprise standard models.

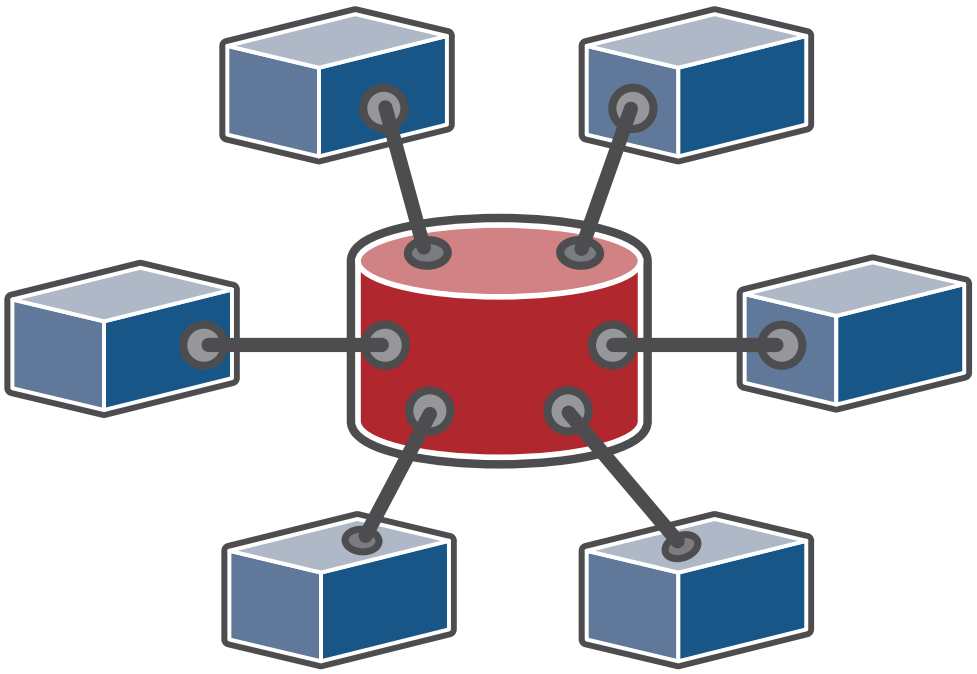


Figure 1: The Next Generation Modular Architecture

Let's consider nine key characteristics of next-generation data integration architectures and how they carefully build on the existing data integration architectures.

1. **Reusability of Integration Logic, Rapid Prototyping, and the Virtual Data Machine**

New data integration and movement technologies appear all the time. The use of virtual business intelligence and Hadoop are relatively new concepts. The ability to reuse integration logic, mappings, and components is a hallmark of next-generation integration environments. Imagine developing in a codeless, graphical mapping environment. You are building a data movement process that shows raw data being moved, cleansed, and combined from multiple data sources into a target warehouse. Now imagine being able to prototype this process virtually without ever moving data from the sources. Then, at the click of a button, without recoding the mappings, you can convert the prototype into ETL and load a traditional data warehouse. This reusability allows fast prototyping of a data warehouse using agile development methods without having to recode as you move from virtual to physical. This same kind of reusability can be used to allow companies to migrate from ETL or extraction, load, and transfer (ELT) or directly to Hadoop, again without recoding the mappings.

This is why next-generation data integration must be built on a virtual data machine (VDM). With a VDM, data integration development is separated from the run-time engine and the run-time engine could be a variety of technologies. The advantage of this approach is that once developers understand the graphical data mapping environment for ETL, they can also do data integration using techniques like ELT, data federation, or Hadoop. The other advantage is that as data changes and grows, different movement technologies may be needed for different integration processes. The VDM approach makes it easy to develop once and deploy anywhere, optimally leveraging the appropriate integration fabric for a given integration project.

2. **Universal Data Access**

The idea of universal data access isn't new. It existed in earlier generations of data integration infrastructure. What is new is the emergence of new and often unstructured data types, business use cases, and platforms such as Hadoop, forcing data integration teams to need to rapidly and effectively leverage enterprise information regardless of its source. Critical to data integration success is the ability to unlock data from such diverse sources as mainframes, custom-built legacy systems, commercial off-the-shelf applications, and Hadoop, while still meeting the speed, quality, and cost objectives of next-generation data integration.

A TDWI survey (Figure 2) clearly reflects the need to access a wide variety of data types, because more than 80 percent of respondents reported that they are either using or anticipate using structured, complex (hierarchical or legacy) real-time message data, spatial data, and unstructured data within the next three years—doubling and tripling usage of the newer data types.[6]

[6]The Data Warehousing Institute, "Next Generation Data Integration," April 2011.

New Data Sources Emerge

For the types of data on the following list, which are you integrating today through your primary data integration implementation? Which do you anticipate using in three years of so?

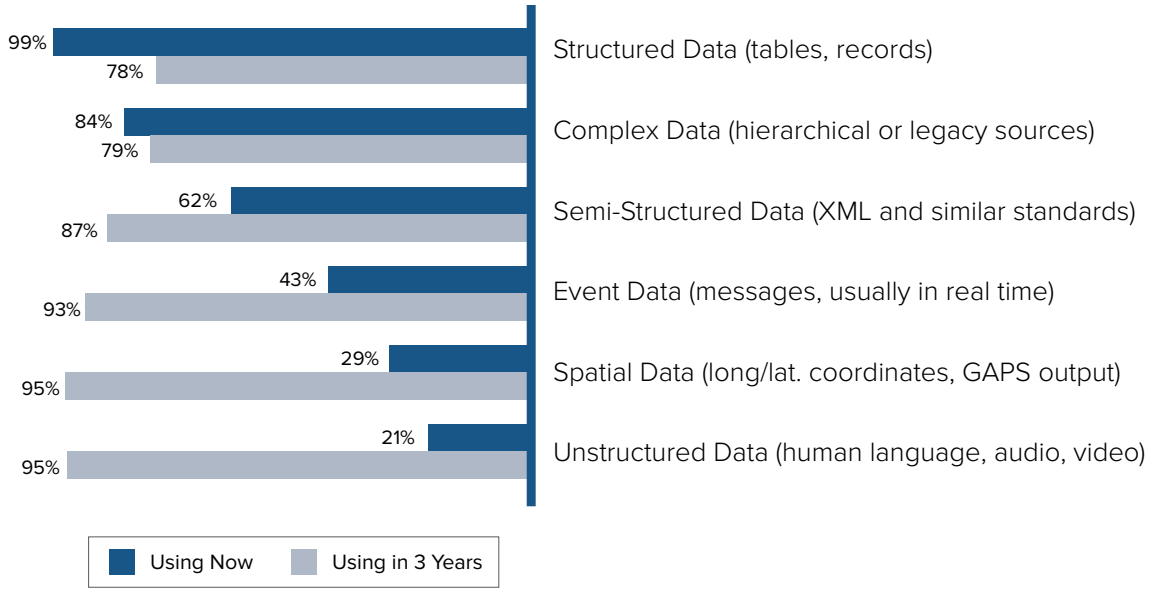


Figure 2: A TDWI study finds strong growth in the use of unconventional data types

3. **Automation of Development, Testing, and Operational Tasks**

With traditional data integration, developers and IT were on their own to deal with many ancillary but critical tasks. Now there is increasing automation around core integration and quality development to make development and deployment processes even more efficient. For example, new technologies allow the enforcement of best practices in developing integration jobs. Imagine having a team of experts who check your integration mappings and identify where you have made mistakes. This kind of intelligence is available today.

In addition, like most software development projects, we estimate that 30 percent of data integration development is spent both testing and validating that integration jobs and the data match the specification. This testing process can now also be automated, reducing at least 50 percent of the time required to QA data integration projects.

Last, much like a manufacturing process, data integration must be monitored to identify problems before they occur. The old approach was to use a generic network and system management platform to perform this task. However, those kinds of systems don't understand the details of the integration process. The current generation of data integration platforms enable very fine-grained monitoring of specific integration processes, step by step as data is extracted, moved, integrated, cleansed, and loaded into target systems.

Next-generation data integration platforms have moved beyond simply providing the blocking and tackling of integration and data quality and are now providing a much greater level of automation that accelerates time to deployment as well as the availability of the overall integration operational system.

4. **Support for Current, Timely Decision Making**

As suggested in the TDWI study, enterprises are looking to expand use of real-time data, with 93 percent of respondents expecting to make use of real-time data within three years. Typical use cases for such data include dynamic pricing, just-in-time inventory, real-time financial transactions, and others related to competitive advantage. Depending on the business case, technologies for data streaming, replication, messaging, complex event processing, virtualization, and change data capture should be considered for reducing data latency and speeding data delivery.

Additionally, in next-generation architectures, data warehouses are beginning to serve as caches to precalculate commonly and frequently accessed data to improve performance and meet service level agreement (SLA) specifications. These precalculations can include complex metrics, fuzzy matching logic, or data cleansing logic that should not need to rerun against the same data, thus slowing performance. Using change data capture technology can help keep these data caches in real-time synchronization with data on the physical systems that can be undergoing changes at any time. Latency of data in multiple caches can be managed to meet use case needs—for instance, zero latency (real-time caching) for fraud detection, minutes of latency for business dashboards, or daily for reporting and ad hoc queries.

Also, while this particular characteristic is notable by itself, you should consider that having real-time data isn't sufficient for next-generation data integration. Key considerations also include the use of a single definition and logical model for both real-time and non-real-time data to minimize rework and the risk of inconsistency, as discussed previously.

5. **Common Metadata Model and Repository**

Metadata management (and a business glossary with semantic data definitions for the business to use and maintain) is still a must-have for next-generation data integration. A common metadata model and repository supplies critical visibility into what can be reused and what is actually being reused. It also enables governance of the integration environment to ensure that roles are following best practices while heightening collaboration between business and IT. By describing data lineage, metadata management provides "factory-floor visibility" that shows where delays exist and where time is being spent, as well as other information vital to continuously improving how IT delivers data projects and initiatives to the business.

6. **Hub-and-Spoke Models**

A hub-and-spoke architecture underlying next-generation data integration meets objectives for greater agility and cost reduction by moving away from traditional point-to-point integration logic that is frequently costly and time-consuming to build and maintain. Promoting reusability with a common platform for "blessed" data, a central hub is particularly important for traditional ETL and MDM initiatives aimed at enabling trusted, timely data on customers, products, suppliers, assets, and more. Respondents to a TDWI survey ranked hub-and-spoke as the second-most preferred architecture for next-generation data integration, trailing only a service-oriented architecture (SOA).[7] In reality, however, an SOA serves as a primary spoke that extends the central hub through services.

[7]The Data Warehousing Institute, "Next Generation Data Integration," April 2011.

BANCO POPULAR CUSTOMER SUCCESS

Banco Popular, a leading financial institution in Spain, Portugal, the Caribbean, and the U.S., has improved the speed, quality, and cost-efficiency of data quality and integration processes used to generate a "quality index" to strengthen regulatory compliance and improve customer satisfaction. The bank uses Informatica® Data Quality and Informatica PowerCenter® to standardize and consolidate millions of customer-related records, replacing a costly and time-consuming process that relied on third parties.

With Informatica, the bank has improved the business value of data and its operational efficiency while reducing risk and cost. "The efficiency, speed, and flexibility that we have gained in the analysis of data are enormous," said Alberto Romero, director of the bank's Quality of Information Office. "We are no longer an office requesting information and instead we have become generators of solutions."

7. **Integrated, Role-based Data Profiling and Quality**

The ad hoc, tactical data quality initiatives typical of first-generation data infrastructures can't meet the ideal of comprehensive, trusted data on an enterprise-wide scale. Effective next-generation data integration seeks to centralize data profiling and quality to improve data visibility and accuracy in support of confident business decisions, while boosting IT productivity and reducing cost and risk. Role-based tools give business analysts a greater stake in identifying and understanding source data and devising and maintaining data quality rules on an ongoing basis in partnership with IT. Importantly, data quality in a next-generation environment is distinguished by its seamless interoperability with data integration technologies, as opposed to standalone solutions of the past.

8. **Enhanced Collaboration and Lifecycle Governance**

Success with next-generation data integration depends on greater collaboration between business and IT. Next-generation platforms must empower business analysts with role-based self-service tools to contribute to data profiling, cleansing, validation, and quality with graphical user interfaces that are understandable to nontechnical users, providing a common ground to narrow the gap between business and IT. Similarly on the IT side, reusability and resource centralization in an ICC can promote cross-team collaboration to speed delivery and improve quality.

A governance framework for managing and streamlining the integration lifecycle helps minimize risk and ensure sustainability through repeatable processes across all aspects of next-generation data integration. Collaboration and governance are particularly important to extending data integration beyond its warehousing roots to encompass MDM, operational data integration, big data analytics, and the introduction of Hadoop and other big data technologies into the enterprise.

In addition to quality, governance frameworks should take into consideration both retention and archiving of data across its lifecycle. As data volumes continue to grow at increasing rates, lifecycle management must be able to seamlessly integrate into the data integration architecture.

The Benefits of Next-Generation Data Integration

Based on the experiences of leading Informatica customers, we can identify distinct improvements through the use of next-generation technologies and techniques, namely:

Streamline Development

According to survey results reported in the 2012 TDWI BI Benchmark Report, changing a data warehouse hierarchy took an average of seven weeks in 2012, up significantly from 4.7 weeks in 2008. Rigid development processes have led to this behavior. Next-generation data integration enables leaders to vastly improve productivity and agility, subsequently reducing costs, through several key steps:

- Prototype the warehouse change using virtualization (aka federation) to ensure the proposed change meets a specific business need, with the business analyst or subject matter expert participating with the IT developer or architect. This should be done before converting the prototype into actual ETL code.
- Use metadata-driven impact analysis to assess the changes, using a micro-project plan, design specification, and test specification, if necessary.
- Replace manual steps by automating testing processes to validate data and data integration processes.
- Use an automated deployment framework to move all required objects from the virtual prototype in the development environment to quality assurance testing to production in the physical data warehouse.

Done correctly, this process can be completed in hours for high-priority requests, in a manner in which the request "pulls" data and it "flows" to the customer, similar to the popular Kanban method of scheduling optimization in lean manufacturing and integration environments.

Assure Confidence

Prevalent data management initiatives (i.e., data integration, data quality, data archiving, data masking, master data management, data warehousing, business intelligence, and analytics) when managed as tactical, IT-driven efforts often deliver solid returns in the targeted environment or business area. But efforts to scale these solutions to support cross-enterprise objectives often fail to meet expectations. To break through this business value ceiling and maximize return on data, you must have confidence in both your data integration processes and the data itself.

While not a sole contributor, technology plays a significant role in helping drive confidence across your organization. Key enabling capabilities include:

- Data quality scorecards to quickly and easily surface quality issues to the business
- Data steward workflows to facilitate the fast and easy resolution of data-related issues and provide comprehensive task workflows that streamline the resolution of data quality problems.
- Proactive data integration monitoring to reduce the risk and increase confidence in data integration processes.
- Common data quality and a master data stewardship to promote broader deployments and increased collaboration across business and IT.

KELLEY BLUE BOOK CUSTOMER SUCCESS

Kelley Blue Book, the leading provider of new and used vehicle pricing values and information at its popular www.kbb.com website, uses Informatica as the central integration technology for its forward-thinking enterprise architecture strategy. Replacing a patchwork of data integration tools, Informatica has given Kelley Blue Book an open, scalable platform for operational data integration and analytics.

Kelley Blue Book uses Informatica to integrate data from scores of internal and external sources—from web clickstreams to vehicle auction data—to optimize the www.kbb.com experience and support critical business initiatives with vehicle dealers and manufacturers. "Using the Informatica Platform, we can process new and existing types of data more efficiently and reduce the time to market for new features for Kelley Blue Book products and services," said Steve Okamoto, director of the Data Services Platform at Kelley Blue Book.

Innovate Without Barriers

Lag times of days, weeks, and even months are not uncommon as IT professionals often start from scratch to access, prepare, and deliver information. This problem is only compounded when organizations look to big data or new computing platforms like Hadoop. The result: frustration for business and IT, the risk of uninformed business decisions, and a general inability to innovate effectively to remain competitive.

Leaders leverage next-generation data integration technology, reusable processes, and a collaborative team to deliver critical information in as little as several hours. Key steps include:

- Work with the business team to identify business potential from unlocking new sources of insight
- Build a roadmap for integration of new sources of data and use of new technologies like Hadoop
- Leverage a virtual data machine approach and separate data integration development from the run-time engine so you can design once and deploy anywhere for utmost flexibility
- Map your program performance by tracking metrics like "first-time through percentage" (e.g., change requests that go into production without iterations) divided by the "time to deliver."

Making Next-Generation Data Integration a Reality

Adopting next-generation data integration is an iterative multiphase effort that depends on clear objectives, systematic measurements, and alignment of business and IT. By focusing initially on discrete, high-value, low-risk projects, your enterprise is positioned for success to breed success, such that value accrues exponentially as agility is heightened across multiple IT and business areas. Some points to consider include:

- **Benchmark your current state.** Identify technological and process weaknesses and quantify the impact of delays, costs, and inefficiencies. Tools like those found at www.governyourdata.com help assess your maturity and identify the business priorities to focus on first.
- **Engage the business.** Find a business champion to support next-generation initiatives; engage the business to pinpoint problem areas and prioritize business goals.
- **Focus on fundamentals.** Devise a roadmap that builds in fundamentals of a next-generation environment, in particular, a strong metadata foundation, data governance, business user self-service, and data quality.
- **Focus on reusability.** Implement the means to harness and expose reusable logic, processes, and resources; explore how an ICC can promote reusability and align stakeholders.

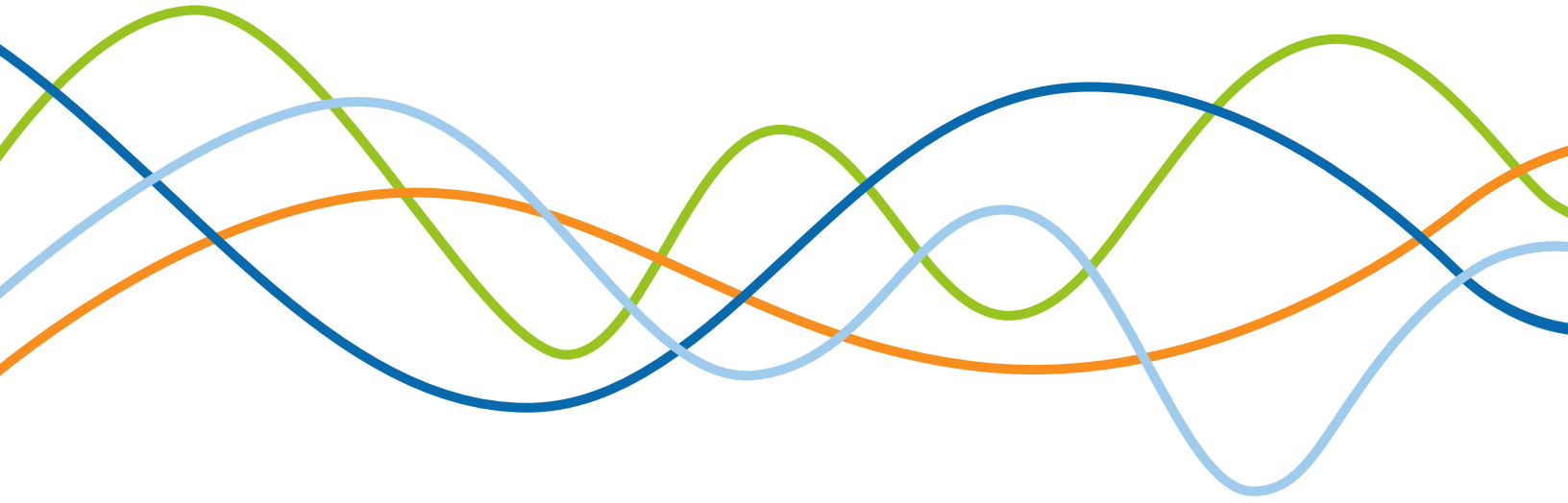
Conclusion

The business and IT pressures that have elevated next-generation data integration to visibility on the enterprise radar screen will only increase in the months and years to come. From big data to globalization and brutal competition to empowered customers, disruptive phenomena are forcing IT leaders to rethink their data integration practices and strategize over new opportunities to turn data into business advantage. In doing so they are able to transform data chaos into breakthrough results through streamlined development, data and process confidence, and a roadmap for data-driven innovation.

An embrace of next-generation data integration concepts, techniques, and technologies is already paying dividends in enterprising organizations across a variety of industries. Next-generation data integration is enabling IT teams to execute no-compromise projects with speed, quality, and cost-efficiency not possible with first-generation approaches. Through a growing body of best practices and technology that continues to evolve with greater capabilities and scope, the strategic objectives of next-generation data integration are within reach of virtually any organization.

ABOUT INFORMATICA

Informatica Corporation (NASDAQ: INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica for maximizing return on data to drive their top business imperatives. Worldwide, over 4,630 enterprises depend on Informatica to fully leverage their information assets residing on-premise, in the Cloud and across social networks.



INFORMATICA®

Worldwide Headquarters, 100 Cardinal Way, Redwood City, CA 94063, USA
phone: 650.385.5000 fax: 650.385.5500 toll-free in the US: 1.800.653.3871
informatica.com linkedin.com/company/informatica twitter.com/InformaticaCorp

IN09_0213_02354