WHITE PAPER

Next-Generation Data Virtualization

Fast and Direct Data Access, More Reuse, and Better Agility and Data Governance for BI, MDM, and SOA

Executive Summary

It's 9:00 a.m. and the CEO of a leading consumer electronics manufacturing company is meeting with the leadership team to discuss disturbing news. Significant errors in judgment resulted in the loss of millions of dollars in the last quarter. Overproduction substantially increased the costs of inventory. The sales team is not selling where market demands exist. Customer shipping errors are on the rise. With competitive threats at an all-time high, these are indeed disturbing developments for the company.

All of these problems can be traced to one thing: the lack of a common view of all enterprise information assets across the company, on demand, when key decision makers need it.

Some lines of business have stale or outdated data. Many define common entities, such as customer, product, sales, and inventory, differently. Others simply have incomplete and inaccurate data, or they can't view the data in the proper context.

A few years ago, a couple of business lines embarked on business intelligence (BI) projects to identify better cross-sell and up-sell opportunities. Their IT teams found that creating a common view of customer, product, sales, and inventory was no easy task. Data wasn't just in the data warehouse—it was everywhere, and it was growing with each passing day. Various tools and integration approaches—such as hand coding, ETL, EAI, ESB, EII—were applied to access, integrate, and deliver all this data to different consuming applications. It became a maintenance nightmare. In the end, the information assets remained within their silos.

Over time, as a result of numerous acquisitions, the IT infrastructure has become even more complex and brittle. Now there is much more data to process and many new data sources to integrate. For example, social media data provides insight into customer sentiment. Legacy data continues to grow and needs to be archived. Other types of data must be leveraged immediately in the context of the data integration.

New projects just add to the complexity. Beyond the original departmental operational BI projects, there are enterprise-wide operational BI, master data management (MDM), and service-oriented architecture (SOA) initiatives.

While critical to the business, a common view of information, available on demand, seems like a pipe dream. What happened? Can the dream become a reality?

This white paper outlines the steps your IT organization needs to take to move its infrastructure from its current state to one that provides a common view and understanding of the business's most important asset—data—and does so on demand. It explains the important role of data virtualization—the process of hiding and handling IT complexity—and how this evolutionary and revolutionary approach can improve the way your IT organization manages and leverages data.

The New IT Hairball

The problem started with what's known as the IT hairball. This hairball took shape as ad hoc and tactical links between any number of applications and databases were created over the years to support any number of data integration requirements. These links, primitive technology such as FTP or other simple file exchange mechanisms, crisscrossed the enterprise, forming a tangled hairball.

This IT hairball prompted the move to more sophisticated integration technologies such as enterprise application integration (EAI) and enterprise service bus (ESB) products. These technologies promised to provide integration mechanisms, align integration strategies, and enable data to be used correctly.

But these integration technologies were used repeatedly throughout the enterprise in tactical, one-off, ad hoc ways. They ended up creating an IT hairball of their own—one with the same complexity and the same problems they originally were supposed to solve. There's a new IT hairball (see Figure 1), and this one is the result of creating multiple layers of data integration technology between the data and its consuming applications.

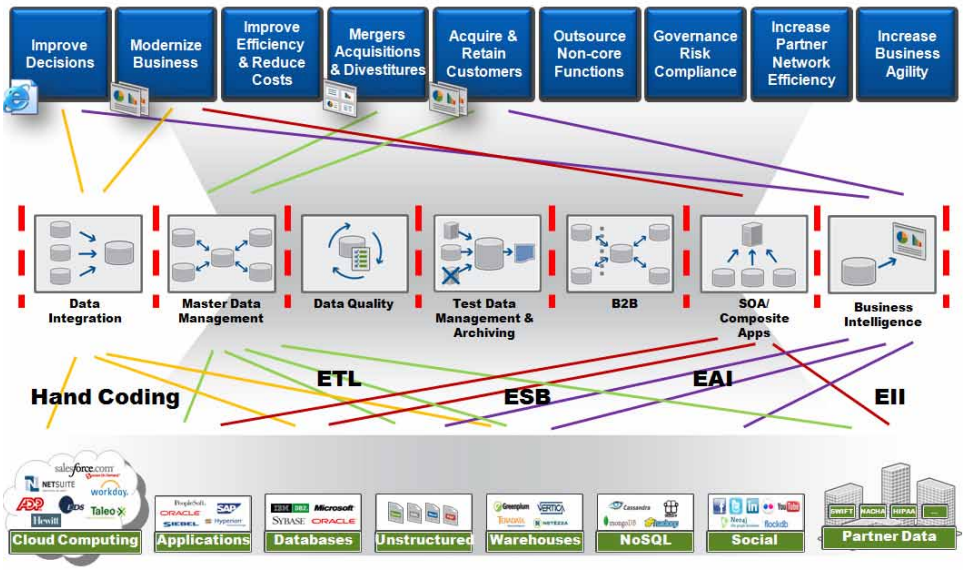


Figure 1: The new IT hairball is the result of creating many layers of data integration technology between the consumers and the physical data.

Today's IT infrastructure can be largely characterized by what it lacks, for example:

- No centralized understanding of where the core data exists in relationship to and in context of other data
- No common way of dealing with the various data semantics used across the enterprise
- No support for applying complex data transformations, such as data quality on federated data, without staging or postprocessing
- No way to involve the business early to identify data issues and define and enforce rules in real time, without any delays or staging
- No real-time or near real-time data access or visibility into data when and how the business needs it.

Additionally, the data integration process is fraught with inefficiencies and delays due to the common business-IT divide. IT organizations must understand both the data integration process and the right enabling technology to bridge the divide and untangle the IT hairball. Fortunately, companies recognize the need for better data integration based on SOA principles for flexibility and reuse (see Figure 2).

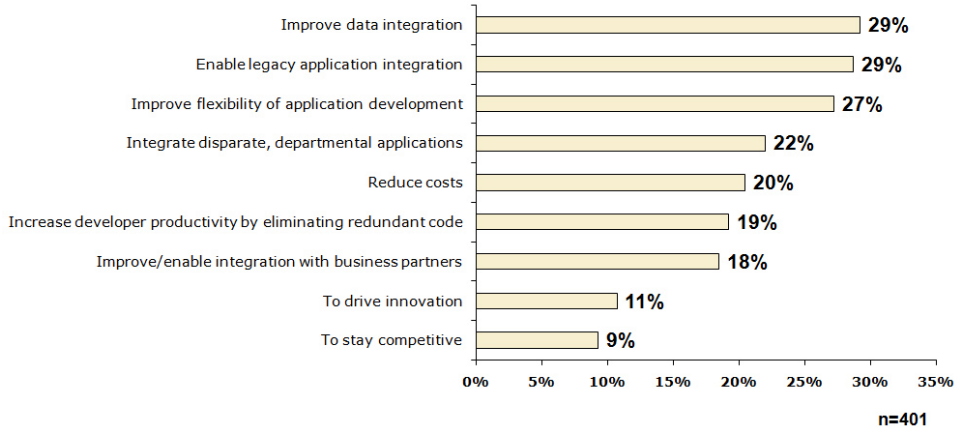


Figure 2: A recent SearchSOA.com survey asked, "What are the top two benefits you hope or hoped to achieve from SOA for your company?" Data integration ties for first place.

The New Solution

IT organizations need a new technology solution to handle the complexities of the new IT hairball. This new solution should complement existing data architectures. It starts with a common data model—a common data access layer that can adapt to the needs of the enterprise. The new solution should be able to:

- Access and federate data (without physically moving it) across all types of data, including operational, analytical, interactional, master, or archived data
- Leverage underlying computing resources and optimizations when possible
- Define and represent data as business entities—such as customer and product—in the way the business thinks about data
- Empower the business to handle some aspects of data integration itself to accelerate data integration, thus avoiding the delays due to IT intervention while still maintaining IT's control over the integration

The new solution should be an adaptive technology that offers appropriate role-based tools that share the same metadata. It should enable the business to collaborate with IT to define and enforce policies around data quality, data freshness, data retention, and data privacy in real time. This common data access layer must then be able to federate across all data and apply rules to the data in real time, without the need for additional processing and staging.

The new solution should bring data virtualization to the enterprise. The underlying complexity involved in integrating enterprise data—the hairball—is both handled and hidden at the same time. Simple, traditional data federation or enterprise information integration (EII) can't deliver this kind of true data virtualization. This isn't simply about federating data in real time. It's about making many data sources look like one and abstracting applications from changes. And more importantly, it's about enabling business and IT to work collaboratively throughout data integration.

The result of this new solution is an IT infrastructure that is agile and flexible enough to deliver the correct data, on demand, in the proper context, to any part of the business that needs it.

The Next Generation of Data Virtualization

Data virtualization is about placing all your diverse physical database assets behind technology that provides a single logical interface to your data. Data is more logically grouped to make that data more useful, which this has been the promise of SOA for years. An SOA with a data services layer is a valid architectural approach to dealing with both underlying business processes as well as sets of services (see Figure 3).

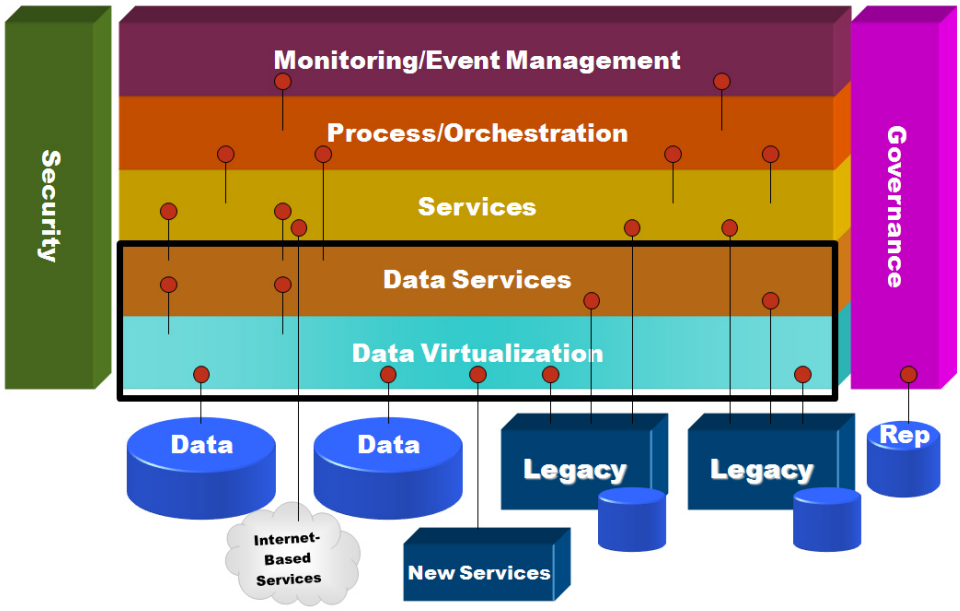


Figure 3: Data virtualization and data services are the foundation of a service-oriented data integration approach

However, you cannot ignore the data integration process that underpins all data integration projects. It all starts with a change request, followed by approval and prioritization of the requests, analysis and design, the creation of prototypes, and then building, testing, and deploying the solution. These steps have stood the test of time and are here to stay.

Yet the data integration process is fraught with inefficiencies and delays owing to errors, missing critical data, data quality issues, the manual back-and-forth between business and IT, and the lack of reuse of data integration logic across projects.

Your data integration solution should include next-generation data virtualization capabilities. Data virtualization based on Lean Integration principles is perhaps the best way to untangle the new IT hairball that limits the on-demand access and use of critical enterprise data. This approach puts the business user—who knows the data best—at the center of the data integration process, while IT maintains the necessary control to govern the process. Business users can define rules and work with IT to enforce rules on the federated data. Role-based tools that share common metadata help accelerate the process and deliver the data, when, where, and how it is needed.

Data integration that includes next-generation data virtualization offers many benefits:

- Fast and direct access to data
- Increased reuse of data integration logic
- Reduced complexity and increased agility
- Better data governance
- Self-service data integration on the part of business users

Let's look at each benefit.

Fast and Direct Access

Providing fast and direct access is at the core of data virtualization. Data is integrated and processed in real time by federating data across several heterogeneous data sources to deliver a virtual view or virtual database. There is no delay due to physical data movement.

When we talk about doing data integration this way, high-performance data processing and universal data connectivity (regardless of the type of data source) becomes very important. Your IT infrastructure must be able connect to all types of data sources, including relational databases, mainframe, unstructured, and cloud. Also, the data virtualization solution that your IT organization selects should provide a high-performance engine that supports caching and all the forms of optimizations people expect from a data federation technology, such as query optimization, rule-based optimization, cost-based optimization, and pushdown optimization.

The data services delivered by a data virtualization solution represent and supply interfaces to both data and behavior using a standard service interface—typically SQL or Web services (Web services that deal with data only or data services). The data virtualization solution must provide an enterprise-grade query engine, query optimizer, and Web services engine. These standards-based data services enable those charged with building business solutions to easily and quickly leverage these services within composite applications, data analysis tools, or anything that needs data access and can consume a service. Templates and data transformation logic can also help expedite the data integration process.

When considering data virtualization, you may also consider the relevance of Big Data. Big Data means all data, including both transaction and interaction data, in sets whose size or complexity exceeds the ability of commonly used technologies to capture, manage, and process at a reasonable cost and timeframe. Big Data includes massive and high-performance databases that leverage distributed approaches to database queries, such as map-reduce. Companies are starting to look at leveraging Big Data for analytical purposes.

Increased Reuse

Increasing reuse has been the battle cry for IT for a while now, but it has largely rung hollow. Data virtualization allows you to place a configuration layer between the data service interface and the physical data itself. This approach greatly improves agility, considering that changes to back-end data sources and the service interfaces are a mere configuration exercise and typically don't require programmatic or deployment changes. The data virtualization layer is a better and more meaningful representation of the business data, which makes it easier to leverage those services from system to system, application to application.

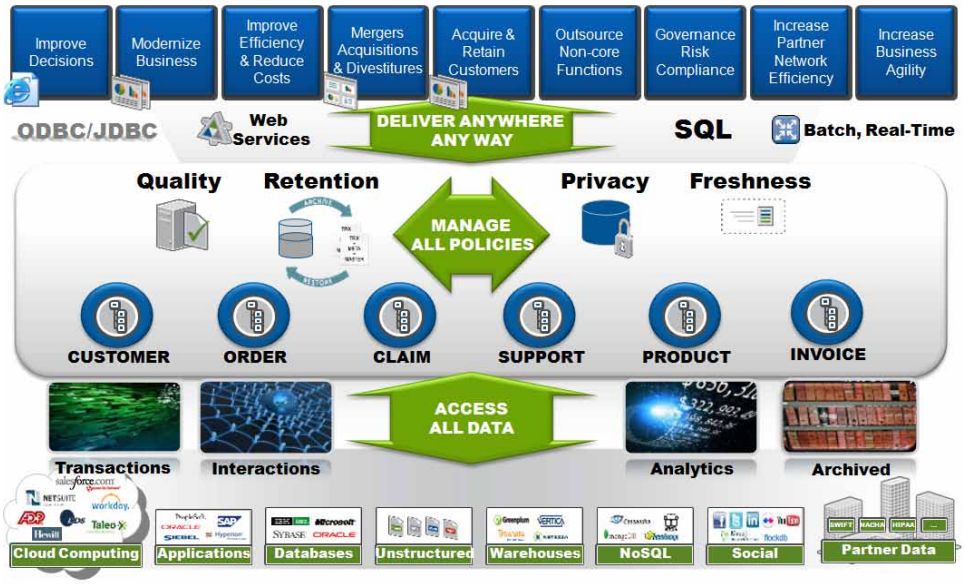
Data virtualization maximizes the reuse of data integration logic across all types of interfaces, such as SQL or Web services, and all modes of data integration, including batch and real time. Data services can be instantly reused across applications by making it easy to access those services, as well as to instantly customize those services for the use case.

Reusing data integration logic means better data security and governance through metadata and data traceability. Data quality—making sure that data is correct in content, form, and currency—can be supported up front in the data integration process. Finally, it is necessary to support the physical materialization of a federated view into a persistent data store such as a data warehouse, if needed at a later stage. this step needs to happen in the same environment; doing so in a separate tool only adds to the complexity.

Reduced Complexity and Improved Agility

Data virtualization's advantage also comes from its ability to reduce complexity. Data virtualization abstracts or hides the complexity of numerous back-end physical data sources and their diverse formats or states (see Figure 4). These data sources are typically created over the years using various integration approaches to solve specific business problems. Over time, many other data sources get added or removed, and information residing within specific applications throughout the enterprise becomes so complex that it's difficult to figure out how key business entities such as customer, sales, and inventory reside within the physical data sources. The complexity makes the core data sources almost unusable.

Figure 4: Data virtualization lets you abstract the underlying data complexities.



Data virtualization reduces the complexity by creating sets of data services or a data abstraction layer that maps from the complex physical data source structures and data, to core business entities, into a virtual database. For example, customer data may exist in 20 or more diverse physical data sources within the enterprise. However, with data virtualization, customer data can be easily accessed, abstracted, and mapped to a single virtual database that provides customer data in a way that the business defines and understands it.

Another important aspect of reducing complexity is to define and apply complex data transformations, such as data quality, in real time to the federated data. Unlike simple, traditional data federation that is typically limited to SQL or XQuery-only data transformations, data virtualization exposes all the rich ETL-like transformations—such as lookups, joiners, and aggregators—which are a prerequisite for enterprise-grade data integration.

Your IT organization should select a data virtualization solution that enables the business to become involved early and often in the data integration process. Such a solution lets the business define rules about the data to address such issues as data quality and data privacy. Because these complex data transformations are applied in real time to the federated data, no time is wasted in staging or additional processing of the data to get it into a consumable state, contributing to the agility of the approach.

Better Data Governance

The business knows the data best. It wants to access and manage data in a way that makes sense to it. However, these needs must be balanced with IT's need to control the data integration process.

Data virtualization can help by enabling better data governance. Data virtualization is a real-time data integration approach. Data is accessed on demand, directly from the data sources, eliminating additional processing or staging of the data. This means that data governance policies defined by the business—such as data security, privacy, quality, freshness, latency, and retirement—must be enforced on the federated data in real time as it flows through the virtual layer. These data policies are defined on the common data model or logical data objects to ensure manageability and reuse.

Simple, traditional data federation technologies assume that the data in the back-end physical data sources are of the greatest quality and in the format the business can readily consume. This is not the way the real world works, as we all know.

Self-Service Data Integration

Traditionally, analysts and stewards who own the data haven't dealt with data integration tasks and technology. Times have changed. Today, self-service data integration places the power of data integration into the hands of those who are closest to and most knowledgeable about data's core business requirements. The benefits of self-service data integration include the ability to drive changes to data integration more quickly and efficiently, and a lower cost of data integration operations, as well as an increase in data service reuse and thus additional value from agility.

However, self-service data integration must be leveraged in the context of data virtualization. The business and IT need to collaborate using role-based tools that share common metadata to find data sources, define the logical data objects, and define and enforce rules on these objects or business entities.

Simplicity is key. Creating data integration flows and services should be a simple configuration approach and not involve programming. For example, the analyst should be able to generate a data integration mapping without facing a great deal of complexity or requiring a deep understanding of the data integration technology.

Informatica Data Services – The Next-Generation Data Virtualization Solution

Informatica® Data Services™ is the next-generation data virtualization solution. As Figure 5 illustrates, Informatica Data Services provides a comprehensive data virtualization best practices framework for:

- Accessing all data
- Defining and applying all policies
- Delivering data to any application, in any way, for all projects

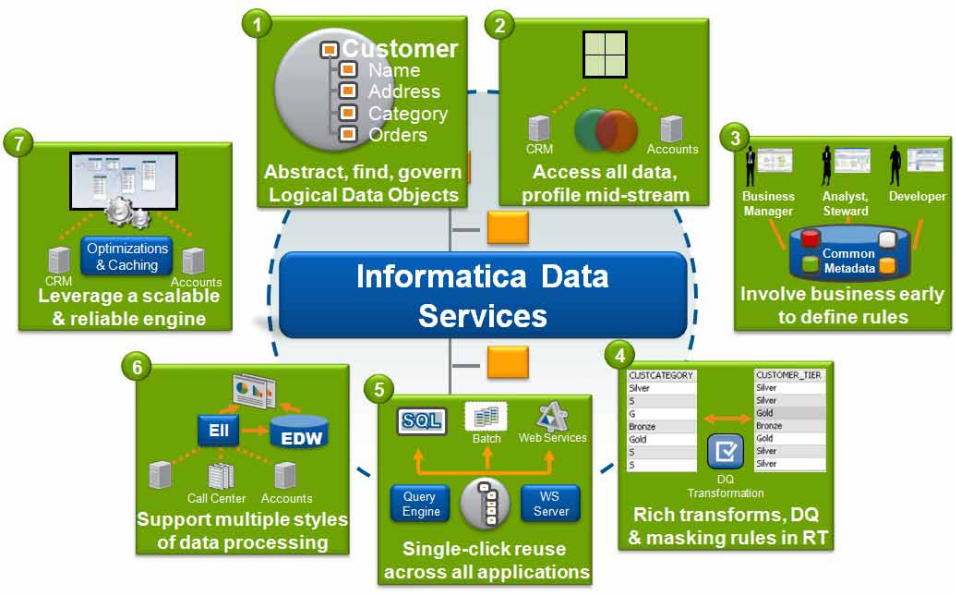


Figure 5: Informatica Data Services delivers a comprehensive data virtualization best practices framework that enables access to all data, defines and applies all policies, and delivers data to any application in any way for all projects

Its robust architecture (see Figure 6) supplies fast and direct access to data of any volume, variety, velocity, or complexity. It enables business and IT users to collaborate using role-based tools to define rules about the data, such as access, freshness, quality, retention, and privacy. It proactively ensures conformance to policies and service-level agreements (SLAs), by enforcing these rules in real time to the federated data. It quickly delivers data services that can be easily reused across all projects such as BI, MDM, and SOA without any rework.

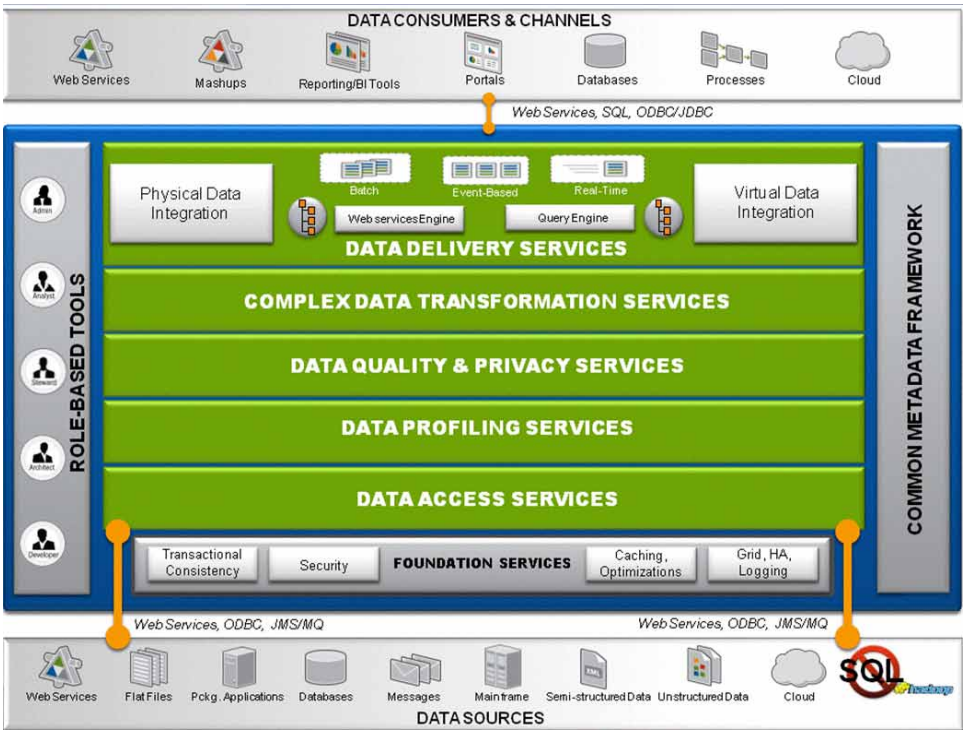


Figure 6: Informatica Data Services' high-level architecture.

Fast and Direct Access to All Data – Analytical, Transactional, Master, Archived, and Interactional Data

With Informatica Data Services, you can provision data services by federating across all types of data in real time and integrate data regardless of whether it is analytical, transactional, master, archived, or interactional data. This data includes map-reduce-based systems, sentiment data from social media, structured data from applications, and unstructured or semistructured data from partners or customers, ensuring your business is not locked out of any data.

Management of All Policies – Access, Freshness, Quality, Retention, and Privacy Policies

You can deploy reusable data services that are compliant with your organization's policies and SLAs for data access, data freshness, data quality, data retention, and data privacy, reducing the risk of noncompliance. Informatica Data Services provides role-based tools for analysts and developers to collaborate on defining and enforcing policies for data access, quality, retention, and privacy dynamically on the federated data, without additional processing or staging.

Support for All Projects – Data Warehousing, Business Intelligence, MDM, and SOA Projects

With Informatica Data Services, you can rapidly and seamlessly reuse data services to accelerate delivery of multiple projects, without any rebuilding or redeployment. These abilities include accelerating data integration projects, operational BI, complementing MDM with transactional data to deliver a complete view, provisioning real-time data quality, exchanging data between partners, and building a data abstraction layer for SOA implementations.

Informatica Data Services provides a comprehensive data virtualization solution that supports a full range of projects and use cases across the enterprise:

- **Data warehouse augmentation**—quickly expand the data warehouse with new data for operational BI and composite applications
- **Data warehouse consolidation**—instantaneously deliver value by leveraging data services to enable reporting against consolidated data across multiple data warehouses
- **MDM hub extension**—rapidly expand the master data in an MDM hub with transactional data to provide a single, complete view of all data
- **Data migration**—seamlessly support data migration projects to insulate reporting users from changes in the underlying data sources
- **Data services for SOA and application integration**—easily deliver a single, reusable data access layer and minimize point-to-point integrations

Conclusion

Data virtualization solves some of the most intractable problems facing your IT organization. Data virtualization places an agile and configurable layer between back-end physical databases and the way these databases are represented using data services.

Informatica Data Services, the industry's next-generation data virtualization technology, hides the complexity of data and gives the power to those who need it to access and use data on their own terms—when they need it, what rules govern the data, how quickly they need it, and how it's represented within the data service. Those who don't want to deal with data's complexity can still get fast and direct access to it. And your IT organization doesn't add yet another layer of complexity to its infrastructure. It offers a graphical, wizard-based, and metadata- and model-driven approach to developing data services that can be instantly reused for SQL querying or Web services interactions. Finally, it delivers all the optimizations required for data federation, including a high-performance SQL query engine and Web services engine, as well as the flexibility to physically materialize federated views when needed.

Informatica Data Services should be considered as your IT organization's data virtualization technology because it:

- Makes multiple heterogeneous data sources appear as one
- Provides a single high-performance environment optimized for both data federation and data integration
- Abstracts data sources from consumers and insulates from change
- Hides and handles the complexity of access, transformation, reuse, and delivery
- Lets the business own the data and define the rules while IT retains control

About the Author

David S. Linthicum

David Linthicum is an internationally known enterprise application integration (EAI), service-oriented architecture (SOA), and cloud computing expert. In his career, he has formed or enhanced many of the ideas behind modern distributed computing, including EAI, B2B application integration, and SOA, approaches and technologies in wide use today.

He is the founder of David S. Linthicum, LLC, a consulting organization dedicated to excellence in SOA product development, SOA implementation, corporate SOA strategy, and cloud computing. He is the former CEO of BRIDGEWERX and former CTO of Mercator Software and has held key technology management roles with a number of organizations, including CTO of SAGA Software, Mobil Oil, EDS, AT&T, and Ernst and Young. Dave is on the board of directors serving Bondmart.com and provides advisory services for several venture capital organizations and key technology companies.

For eight years, Dave was an associate professor of computer science. He continues to lecture at major technical colleges and universities, including the University of Virginia, Arizona State University, and the University of Wisconsin. Dave keynotes at many leading technology conferences on application integration, SOA, Web 2.0, cloud computing, and enterprise architecture and has appeared on a number of TV and radio shows as a computing expert.

Learn More

Learn more about Informatica Data Services and the entire Informatica Platform. Visit us at www.informatica.com or call +1 650-385-5000 (1-800-653-3871 in the U.S.).

© 2011 Netspective Communications LLC

1602 (05/24/2011)