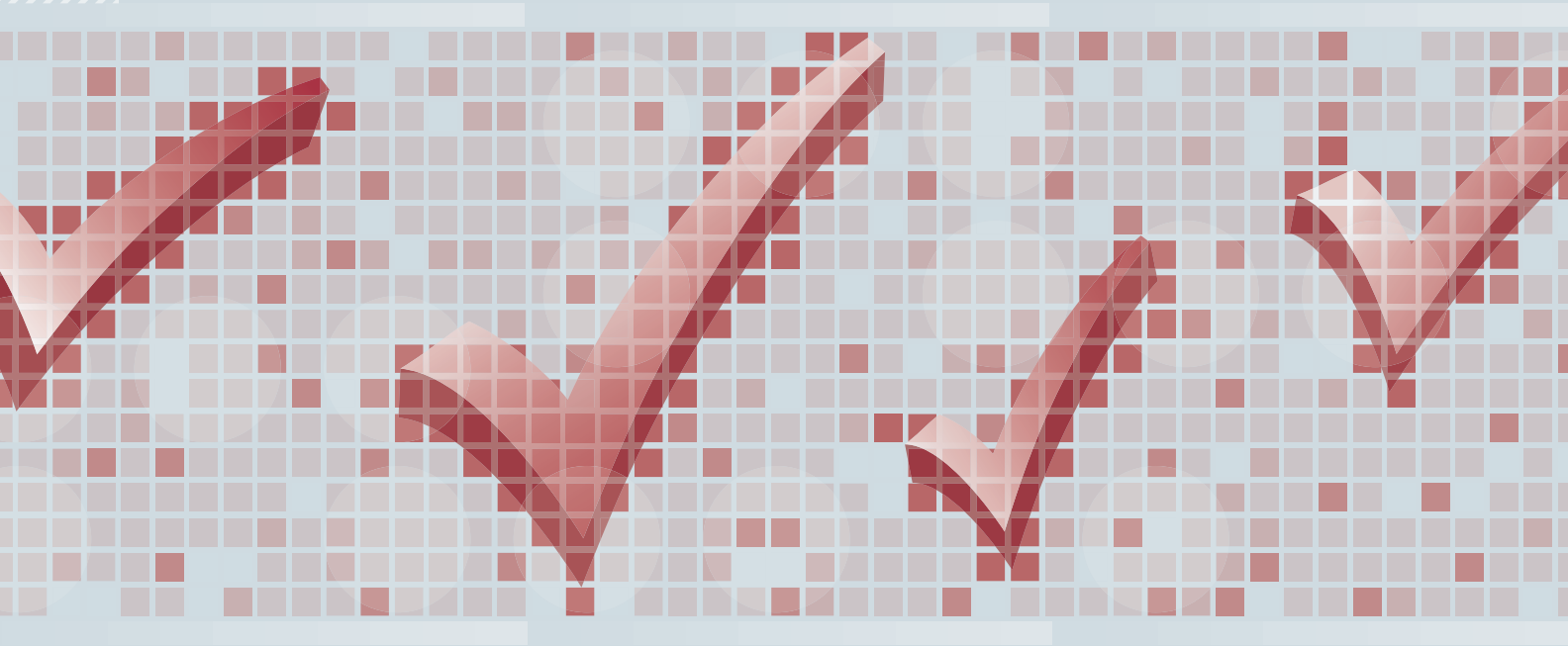
TDWI RESEARCH

TDWI CHECKLIST REPORT

# TAKING IT TO THE NEXT LEVEL: NEXT-GENERATION DATA INTEGRATION

By David Loshin



Sponsored by

INFORMATICA®

tdwi.org

tdwi

MARCH 2013

TDWI CHECKLIST REPORT

# TAKING IT TO THE NEXT LEVEL: NEXT-GENERATION DATA INTEGRATION

By David Loshin

## TABLE OF CONTENTS




1201 Monster Road SW, Suite 250
Renton, WA 98057

**T** 425.277.9126
**F** 425.687.2842
**E** info@tdwi.org

**tdwi.org**

## FOREWORD

The information technology (IT) vision for the future lies securely in the realm of data information. This we have learned from the growing publicity of "big data," analytics, and the promise of improving the customer experience while increasing revenue via the collection, analysis, and synthesis of structured and unstructured data. Visionary stakeholders have recognized the value of expanding the visibility of shared information across business functions.

As data growth rates are predicted to increase five- to ten-fold by 2015, there is clearly a point at which the cost and effort associated with managing data will far outpace the rate of data volume expansion. And despite the decreasing cost of commodity hardware and availability of COTS tools for data integration and analytics, companies continue to struggle to deliver comprehensive data visibility in a way that meets business needs within a reasonable time frame, while also containing the potential for increasing costs and fostering innovation.

Yet there are organizations that have succeeded at lowering costs and reducing delivery times while improving overall information visibility and usage—while also paving the way for leveraging emerging technologies such as the cloud or Hadoop. These organizations are the ones that have designed an IT road map that accommodates the expanding dependence on data integration and information availability through improved data management productivity and agility. This IT road map also transforms data integration from a series of tactical IT projects into an enterprise business function that is easily adaptable for future business demands.

The next generation of data integration takes this into account by seamlessly combining streamlined development with fully integrated data accessibility, quality, and utility. This TDWI Checklist Report presents a series of recommendations for developing a value justification for engaging business sponsors, as well as developing a strategic IT road map promoting the value of data architecture and data integration. In turn, informed decisions for assessing enterprisewide requirements for data acquisition, intake, integration, sharing, and retention will help companies succeed in satisfying the expanding breadth of the user communities, the growing desire to analyze rapid streams containing massive data volumes, and the increased reliance of automated systems on analytic results. These recommendations will help guide the reader in making decisions to manage the need for productive, yet innovative, data integration projects while simultaneously managing effort and costs with the right amount of oversight.

### NUMBER ONE

ENGAGE WITH BUSINESS SPONSORS TO CLARIFY INFORMATION VALUE AND PRIORITIZE DATA INTEGRATION INITIATIVES.

Key organizational stakeholders are finally recognizing that there is inherent value in information, despite years of application development that largely centered on satisfying acute functional requirements while ignoring the significance of the data acquired, created, and used. The challenge lies in effectively translating the perception of "data as an asset" into concrete terms relating the costs of data management activities to the business value created as a result, either through the creation of new opportunities or by addressing critical operational problems. This is particularly critical as IT and data management proposals draw increased scrutiny; a clear value proposition is the best way to have funds budgeted and allocated to pay for data integration projects.

Yet the concept of quantifying the value of data as an asset is driven by considering data and information as "first-class citizens." Transition the theoretical concept of "data as an asset" into something that is not only actionable but will also frame the cost/benefit analysis for ongoing and future data integration tasks. Develop templates, models, and processes for valuation of information as a corporate asset by linking business value drivers such as increased revenue, decreased costs, and reduced risk to data management and data integration activities.

Although data integration is currently seen as a collection of distinct technical activities that are broken up across and outside the organization, the future of creating value from information is tightly bound to the ability to move and integrate data from an end-to-end, intake/creation-to-delivery perspective. A value proposition for data integration can be used to engage business sponsors, clarify the business need for specific data integration tasks, and then help identify and prioritize data integration initiatives. This means focusing attention on data integration competency as a success factor for the future, and considering the holistic nature of the different types of data integration capabilities as they extend beyond the ETL pigeonhole to encompass data intake, organization, transformation, replication, coherence, and delivery (among others).

## NUMBER TWO

DEVELOP A STRATEGIC ROAD MAP AND PROGRAM FOR DATA INTEGRATION AS A DISCIPLINE.

If information is to be considered a critical business asset, then all aspects of its management warrant proper engagement from the user community in establishing standard practices and processes that lead to scalable (albeit flexible) methods that can be deployed consistently across the enterprise to ensure a level of trust. Developing a strategic plan aligns the acquisition of tools and technology with the continuously developing demands of both production and emerging business applications. This strategic plan must map current and future needs to an inventory of standardized practices that make the best use of data integration tools and automation. Some steps in developing that strategy are to:

- Understand both the current and future end-user information usage scenarios and document the dependence on enterprise data sets
- Document the end-user availability, access, currency, quality, scalability, and system performance requirements
- Map the information flow from intake/creation to point of use to locate use of data integration techniques
- Assess the current state of data integration capabilities
- Analyze gaps in the way the current state satisfies the existing and future requirements
- Identify the data integration capabilities (such as managed data access, extraction and transformation, replication, data virtualization, low-latency messaging, changed data capture, etc.) necessary to satisfy unmet requirements
- Redesign the data integration architecture based on the integration of required technical capabilities and corresponding levels of maturity
- Develop a program plan for acquiring technology, training and skills acquisition, and incorporation of best practices for data integration into the enterprise system development life cycle

When this approach is applied across the organization and the requirements are evaluated from an enterprise perspective, the strategic road map will unify the aspects of data integration as a holistic discipline with its own capability/maturity model. At the same time, performance measures can link the availability, access, currency, quality, scalability, and system performance requirements to key business value drivers. Establishing this context for linking data integration capabilities to business value provides the practitioners with a basis for engaging business users and securing their continued support, based on their identified needs, in ways that are in step with the corporate value proposition.

## NUMBER THREE

EVOLVE THE DATA ARCHITECTURE TO SUPPORT KEY EMERGING TECHNOLOGIES.

Your strategic data integration road map will likely result in guidelines for redesigning the architecture for data management and, correspondingly, for data integration, especially in the face of exploding data volumes and widening variances due to the peculiarities of unstructured data, all at increasing speed. A benefit of redesigning the data integration and data management architectures is that standard approaches for metadata-driven, rules-based processing can be easily adapted to support many data integration processes supporting a variety of business process initiatives.

Design an architecture that supports the data integration road map and the key aspects of information delivery: intake of massive amounts of data, data transformation and organization, data validation and quality assurance, self-service access, collaboration, testing and operations automation, and other best practices to help improve overall data integration productivity. Look beyond the conventional approaches to data warehousing when considering ways to support the breadth of application integration use cases. Improving overall business visibility means you will eventually need to manage (and govern) data for all integration projects to ensure data consistency between sources and targets, as well as coherence of information among the different data consumers.

Ensuring that standard approaches can be put in place depends on streamlining the use of automation tools. Employing a variety of non-interoperable tools from different vendors will complicate the transition within existing budgets and time frames. First, consider the types of standards for metadata, business rules, and information exchange, and specify that compliance with those standards is a benchmark for data integration interoperability. Second, make sure that product road maps are aligned with the trajectory for emerging technologies.

In other words, choose vendors that not only have a historical track record in supporting end-to-end data integration needs, but also currently support or plan to support evolving information accessibility demands, including real-time data integration, collaborative business glossaries and metadata management, fully integrated data quality, alignment with master data management, connectors to “utility-style” cloud computing, and big data methods and frameworks such as Hadoop, as these are the most common technologies needed for future projects.

## NUMBER FOUR

ADJUST THE DATA INTEGRATION AND ANALYTICS ARCHITECTURE TO SCALE WITH BUSINESS PRODUCTIVITY DEMANDS.

The traditional approach to addressing system performance gaps is to throw more hardware at the problem. Increased investment in hardware capital assets may temporarily alleviate the system performance gap. However, given an annual average data volume growth rate of 50–60 percent (and that is just for transaction-related data!), increased capital expenditures for hardware will only drive additional costs, effectively resulting in an ever-increasing "annual premium." Yet the demand for mixed use of analytics delivered to a broadened user community means that the driving force of business productivity relies on the ability to scale the environment with nimbleness and agility.

A more cost-effective approach is to reconsider expensive hardware expenditures when given the opportunity to redesign the data management architecture. Evaluating alternative performance frameworks and methods allows you to take advantage of evolving approaches to performance computing. For example, the elasticity of big data and grid platforms based on commodity hardware can be used to satisfy a mixed-use environment for reporting, ad hoc queries, and intense algorithmic analytics.

At the same time, one can begin to offload data integration and preprocessing from expensive platforms (such as specialty hardware appliances) to commodity-based environments that can easily scale linearly with the data volume growth curve. Introducing new techniques that employ best practices, such as data replication and changed data capture, are effective at leveraging commodity hardware by moving the query loads for expensive applications from specialty hardware to lower-cost data stores.

And in reaction to the ever-growing mass of data being created and acquired, it is valuable to more aggressively manage organizational data retention policies. Despite the users' desire to "keep everything," there is a lot of data on source systems and data warehouses that is no longer needed. Archiving this data alone has given some companies years of future room for growth. This is a good short-term initiative because the money saved by not spending hundreds of thousands of dollars (if not millions) on new disk storage can instead be invested in alternate technologies that support the longer-term architecture design requirements.

## NUMBER FIVE

EMPLOY DATA INTEGRATION AND METADATA TOOLS FOR DRIVING CHANGE.

Absorbing different kinds of acquired or created data sets originating from a variety of structured and unstructured sources poses one of the biggest challenges in data integration. This is complicated by different definitions associated with common data element names within structured data, as well as subtle (or sometimes blatantly obvious) contextual changes in meaning ascribed to business terminologies parsed and extracted from unstructured data. Making sense of the accumulated data sets requires defined business term glossaries, concept hierarchies, and business rules, as well as agility as changes appear in the data that impact business processes downstream. The future vision for data integration must inherently establish confidence in the integration processes, with continuous monitoring and reporting of the processes to ensure observance of defined business rules and policies.

Your organization can benefit from managing known business terms, along with a comprehensive map of the ways those terms are defined and how they are mapped to specific entities and data elements used across the enterprise. Because increased cross-business function data interrelationships create more information dependencies, the need to identify potential semantic conflicts becomes that much more acute.

Your organization needs to manage the metadata, business glossaries, and data lineage related to the observance of business policies to track what the business needs to deliver while simultaneously determining the impacts of any changes. Collaborative processes and tools for managing shared metadata can be used to capture the different business terms and corresponding definitions, harmonize them when possible, and differentiate them when necessary. In turn, these metadata artifacts can inform the development of common data integration patterns and models.

At the same time, these types of collaborative, shared metadata management platforms can facilitate the necessary flexibility to continue existing operations in the presence of change, while nimbly adapting to emerging business realities. As business term usage evolves, the metadata management platform can be used to evaluate differences in meaning, and ultimately assess impacts of change to dependent business applications. This level of visibility helps the project planner scope any effort for modifications to existing systems to address those changes.

## NUMBER SIX

DEVELOP TOOLS FOR SELF SERVICE AND COLLABORATION, AND ENCOURAGE IT.

One of the most frustrating aspects of a business user's dependence on the IT team for supporting the development of reports is the delay in clarifying requirements, which results in an even longer delay in delivering the expected data. These delays prevent business users from identifying and consequently exploiting emerging opportunities. An effective way to alleviate this bottleneck is to hand some degree of control back to the users via a parameterized self-service mechanism.

Providing a self-service capability for the user eliminates two productivity bottlenecks. The first bottleneck is caused by the iterative interactions to solicit, document, and review user requirements, while the second involves the time to prototype and configure the delivery of the requested information to the end users. Instead, self-service capabilities in data integration provide rapid prototyping tools, specifically designed for analysts, that give them direct access to data inside and outside warehouses.

Introduce the appropriate data integration methods, tools, and layered services to enable self-service data access that can be directly executed by the business users. Reducing the turnaround time for satisfying user information access needs will reduce operational costs while potentially opening new revenue-generation opportunities. Yet while the business payback is in months, introducing self-service data access will require commitment from the business because you are changing part of the development process. Therefore, be sure that the business is committed to supporting the process and that the right verification and validation controls are in place to be successful.

## NUMBER SEVEN

PROMOTE THE PROPER LEVEL OF DATA GOVERNANCE.

Over the past few years, the conventional wisdom is that introducing a level of oversight for data management will lead to overall improvements in end-to-end information processing. However, attempting to deploy the types of full-fledged data governance frameworks depicted in white papers and theoretical courses often results in an organizational structure and hierarchy for creating enterprise data policies that lacks the authority, power, or training to operationalize or ensure observance of those policies. Participation in these heavy-handed data governance organizations often peters out when the effort and costs seem to outweigh the observed benefits, and when observance of data policies is perceived to be a barrier to progress.

Instead of a heavy bureaucracy destined to mill reams of unobserved data governance policies, consider the introduction of a limited data governance administration that does not prevent progress from being made. As part of the data governance activity, clarify the responsibilities of the members of a data governance committee as well as the role of the data stewards in deploying and ensuring observance of data policies. Concentrate on the practical aspects of governance and limit the scope of data policies to those that are directly linked to business value, can be implemented in ways that result in quantifiable measures, and can be naturally embedded as a set of services supporting the data integration discipline. Select the right opportunities to introduce data governance practices as part of your project road map by introducing the practices needed to make a project successful. Incremental establishment of governance with clear milestones and deliverables will demonstrate value over time in ways to best take advantage of funding opportunities as they arise. This will allow data governance to grow progressively over time.

## AFTERWORD

Looking forward, a full complement of capabilities supporting a data integration discipline is destined to rapidly become the keystone of the future IT vision as a way to control data management costs while satisfying business user needs. In summary, this Checklist Report promotes the development of an information architecture that leverages best practices in data integration under the direction of a properly scoped data governance framework.

Define data policies that can be incorporated into the end-to-end data integration fabric, and adopt policies that can accommodate data scalability while allowing for flexibility into future design and enhancements. Promote opportunities for enabling end users in controlled ways that balance observance of data policies while reducing the number of data requests that constrict the IT bottleneck.

Finally, assess the complete set of requirements for data integration capabilities and map out a strategy for introducing these capabilities in lockstep with business needs. Consider vendors whose tools best satisfy the data integration requirements, with interoperable components connected via "smart" metadata to drive productivity, agility, and confidence, and whose product road map is aligned with key emerging technologies.

## ABOUT OUR SPONSOR

INFORMATICA®

**www.informatica.com**

Informatica Corporation 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 5,000 enterprises depend on Informatica to fully leverage their information assets residing on-premises, in the cloud, and across social networks.

## ABOUT THE AUTHOR

**David Loshin**, president of Knowledge Integrity, Inc. (www.knowledge-integrity.com), is a recognized thought leader, TDWI instructor, and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices, as the author of numerous books and papers on data management, including *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is a frequent invited speaker at conferences, Web seminars, and sponsored websites and channels, including www.b-eye-network.com. His best-selling book, *Master Data Management*, has been endorsed by data management industry leaders, and his valuable MDM insights can be reviewed at www.mdmbook.com. David can be reached at loshin@knowledge-integrity.com.

## ABOUT THE TDWI CHECKLIST REPORT SERIES

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

## ABOUT TDWI RESEARCH

TDWI Research provides research and advice for business intelligence and data warehousing professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence and data warehousing solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.