TDWI RESEARCH

TDWI CHECKLIST REPORT

Modernizing Data Integration

to Accommodate New Big Data and New Business Requirements

By Philip Russom

Sponsored by:

informatica sas

tdwi.org

tdwi
Advancing all things data.

DECEMBER 2015

TDWI CHECKLIST REPORT

Modernizing Data Integration

to Accommodate New Big Data and New Business Requirements

By Philip Russom

tdwi
Advancing all things data.

555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

TABLE OF CONTENTS

FOREWORD

All of us in data management are experiencing an extended period of great change as big data, other categories of new data, and new data management platforms enter our organizations. In response, most user organizations are scrambling to learn the new technologies and—more important—how to leverage the new data and platforms for business advantage. As a result, many data professionals are now facing new requirements as well as future requirements that will arise as new data sources come online.

The changes afoot are driving many technical organizations to rethink and modernize their data management infrastructure, team, and skills. Among these efforts, *data integration modernization* is one of the most pressing because of the broad role data integration (DI) plays in capturing, processing, and moving data—both old and new. Without modern DI solutions, organizations cannot satisfy new and future requirements for big data, analytics, and real-time operation.

DI modernization can take many forms depending on the current state of your infrastructure and what kinds of new data or platforms you must embrace. Instead of trying to list them all here, we offer recommendations for the seven most pressing DI modernization tasks. These recommendations can guide your modernization efforts as you select vendor products and update your solution designs:

1. Complement the high latency of older DI practices with a broader range of data ingestion techniques
2. Embrace the new data prep practices and tools for agility, speed, simplicity, and ease of use
3. Integrate data in ways that enable self-service access to new and big data for a wide range of users
4. Modernize your data integration infrastructure by leveraging new data platform types
5. Keep adding more right-time functions as you modernize your data integration solutions
6. Modernize your data integration functionality to finally get business and analytic value from multi-structured and non-traditional data
7. Consider modernizing your DI tool portfolio with an integrated platform of multiple data management tools

This TDWI Checklist Report drills into each of the seven recommendations and discusses many of the new vendor product types, their functionality, and user best practices that can contribute to DI modernization. We also present the business case and technology strengths of each recommendation.

☑ NUMBER ONE

COMPLEMENT THE HIGH LATENCY OF OLDER DI PRACTICES WITH A BROADER RANGE OF DATA INGESTION TECHNIQUES

One of the most dramatic changes in data integration practice in recent years is the modernization of data ingestion. Ingestion is simply how, where, and how frequently data entering an environment is landed or loaded into targets (such as data staging areas, data warehouses, or file systems). For decades, the ingestion processes of ETL-style data integration practices have been "latent" (i.e., time consuming, often running overnight). TDWI survey data shows that most data in most data warehouses is refreshed on a 24-hour cycle. However, the percentage of data that is quickly or frequently collected, prepared, and delivered for presentation continues to increase for several reasons:

Some new sources of data generate data frequently. An important category of big data is machine data. Sensors are built into (or added to) a growing list of machines including vehicles, hand-held devices, and production-line robots. Machines aside, GPS sensors are proliferating on shipping pallets and other mobile assets. Some sensors generate and broadcast data in a continuous stream of events; others broadcast only when pinged (as with RFID chips) or when the sensors' machine takes action (e.g., when a robot installs a widget). The point is that many organizations want to capture and leverage new streaming sources to improve logistics, sentiment monitoring, service-level agreements and quota compliance, facility surveillance, operational analytics, and business-activity monitoring.

Business practices that require very recent data continue to grow. Users have long been practicing operational business intelligence (OpBI), which frequently updates management dashboards and other operational reports with data that ranges in freshness from several minutes to a few hours. Similar to OpBI, practices in performance management, reporting, OLAP, and advanced analytics demand fresher and fresher data. Fresh data gives these businesses a greater competitive edge, enhances customer relationships, improves operational excellence, and enables nimble but informed tactical decisions.

Data ingestion practices need to accommodate data of many speeds and frequencies. Don't forget: you still need latent ETL and ELT to maintain accuracy and for the extreme joins, transforms, and audit trail typical of data for data warehousing, most standard reports, and many OLAP solutions. The challenge is to design new data integration (DI) solutions (or adjust older ones) to capture and ingest new data faster and more frequently. This is sometimes called early ingestion or continuous ingestion, which is quite fast and frequent compared to overnight batch loads. As a trade-off, early ingestion does little or no transformation or aggregation of data prior to load because that would slow down ingestion. A benefit is that the data is captured in its original state, which means it can be repurposed repeatedly as new requirements for reporting and analytics arise. The greatest benefit is that the data is ready as soon as possible for reporting, analytics, and operations.

Again, a slight challenge with the approach is that repurposing data is increasingly performed on the fly at runtime (instead of prior to load time), as when a data analyst or data scientist explores data and develops new data sets for analytics. As another example, a DI routine may parse changed data (newly ingested) to update intraday operational reports or analyses. A sales manager may refresh a cube to look at today's sales so far.

Today's modern hardware and software are fast and scalable, so continuous-ingestion performance and runtime processing are now practical. In addition, tool functionality for time-sensitive data processing has reached a new level of maturity, as seen in functions that this report will explain, such as event stream processing, data federation, self-service data access, and data prep.

Data quickly ingested may also be processed in traditional ways. Continuous ingestion makes new data available to technologies and users needing it immediately, whereas the "operationalization" of captured data later calls on established best practices in data quality, modeling, and aggregation. For example, the online processing of stream data from manufacturing robots can reveal bad lots or other material problems that need immediate attention. The same data studied offline reveals equally valuable long-term trends in supplier performance relative to product quality.

☑ NUMBER TWO

EMBRACE THE NEW DATA PREP PRACTICES AND TOOLS FOR AGILITY, SPEED, SIMPLICITY, AND EASE OF USE

As mentioned, processing data on the fly has emerged as a distinct data integration practice. The practice has many names including data wrangling, data munging, and data blending. Some people call it "DI light" because its implementations are usually a small subset of DI functionality, trimmed for reasons of usability and performance. However, the name most often heard at TDWI is "data prep," short for *data preparation*. A number of tool types support some form of data prep, including those for data integration, data profiling, data quality, data exploration, analytics, and data visualization.

In particular, data prep is now common for many forms of analytics. It enables a data analyst, data scientist, or similar user to work with detailed source data (full of rich details) without being hamstrung by existing data models and standardizations. After all, this kind of analytics is typically a discovery mission, and preparing data stringently (as ETL for data warehousing [DW] does) can remove the valuable nuggets that an analyst is trying to discover, such as outliers that suggest a new customer segment or non-standard data that suggests fraud or unauthorized access.

Analytics aside, data prep and data exploration often go hand in hand, as when a user explores large collections of data, typically those managed in data lakes, data vaults, enterprise data hubs, and some data warehouses. The user builds a data set as s/he explores, and that data set is then used for data analysis or visualization. In a related example, data prep often combines with functions for self-service data access and self-service report creation or analysis, as discussed in the next section of this report.

Note that data prep is complementary to traditional data management practices. The two apply to different users, apps, and other contexts. In general, new data prep is typically for data exploration and analytics, not permanent designs or highly accurate reports. The two can work together: data sets first constructed via data prep (in support of data exploration and analytic practices) can become a prototype for permanent data sets when the outcome of exploration or analysis is operationalized. During operationalization, the output of data prep is greatly enhanced and improved, using functions for data quality, transformation, modeling, and aggregation. Hence, the modern DI portfolio should include tools that support both practices.

Speaking of tools, data prep regularly taps data federation and virtualization functions. These are ideal for table joins, light transformations, and access to multiple data platforms usually required of data prep. Data federation and virtualization create dynamic, integrated views of disparate data that enable data prep to operate virtually.

☑ NUMBER THREE

INTEGRATE DATA IN WAYS THAT ENABLE SELF-SERVICE ACCESS TO NEW AND BIG DATA FOR A WIDE RANGE OF USERS

A growing class of end users is making greater use of self-service functions in a wide range of software tools and platforms including tools for reporting, analytics, and data integration. Self-service functions are strewn across diverse tools because the users themselves are diverse, ranging from highly technical personnel (data analysts, data scientists, and other data management professionals) to mildly technical business people (data stewards, business analysts, and other power users). Due to the diversity, self-service takes multiple forms including self-service data access, data prep, report creation, visualization, and analytics.

Self-service data functions are important. They enable users to work with data with spontaneity, speed, and agility because users aren't waiting for IT or a data management team to create a unique data set, report, or analysis for them. IT and other teams, in turn, are off-loaded when self-service data is set up so users can create their own data sets and the reports and analyses based on them. According to a recent TDWI report, the four tasks BI users want to do most via self-service are (in priority order) data discovery, visualization, dashboard authoring, and data prep. This is more than wishful thinking; the same report reveals that half of users are already practicing data-driven self-service successfully.[1]

Data integration modernization should enhance self-service data. Enterprises have great interest in supporting more and better self-service data access and data prep, as well as users performing data exploration, reporting, and analytics. This can be achieved in different ways:

- **Integrate data specifically for self-service.** For years, data warehouses and marts fulfilled this requirement. However, warehouses and marts are mostly aggregated and calculated values. These are still relevant, although the trend is toward detailed source data collected in databases and file systems. Growing practices such as data exploration and advanced analytics work well with raw, untouched data. In response, modern data integration solutions nowadays feed new and big data into data lakes, data vaults, and enterprise data hubs that may be housed on Hadoop ecosystems, relational databases, or file systems. Note that these complement (but don't replace) traditional warehouses, marts, and cubes.
- **Depend on self-service tool functions.** In data integration tools, functions involve high ease of use and business-friendly views of data, which can enable self-service data access and data prep. Though designed for less technical users, TDWI also sees highly technical users availing themselves of these functions because everyone benefits from their agility and autonomy. Note that even when ease of use is high, less technical users still need training in both the tool and the best practices of data management.

[1] See Figure 9 in the 2015 TDWI Best Practices Report *Emerging Technologies For Business Intelligence, Analytics, and Data Warehousing*. Available as a free download at www.tdwi.org/bpreports.

☑ NUMBER FOUR

MODERNIZE YOUR DATA INTEGRATION INFRASTRUCTURE BY LEVERAGING NEW DATA PLATFORM TYPES

One of the most exciting developments in recent years for data professionals is the arrival of several new data platforms, such as the Hadoop family of open source products and new database management systems (columnar DBMSs, appliances, graph databases, and NoSQL). Most of these are available on premises or in the cloud, showing that cloud and SaaS are now important components of the infrastructure for DI, DBMSs, and other data platforms. Although these are DBMSs or other types of data platforms (note that Hadoop is not a DBMS), all have positive ramifications for modernizing your DI infrastructure.

For example, a Hadoop cluster has several roles to play in modernizing DI, especially when DI supports a multi-platform data warehouse environment (DWE), as in the following examples:

- **Hadoop is an effective data landing area for many feed speeds and data types.** The Hadoop Distributed File System (HDFS) is suited to high-latency batch, low-latency microbatch, stream capture, and continuous ingestion. Furthermore, file-based HDFS can capture, manage, and process any data that can be stored in a file.

- **Hadoop is a scalable and powerful data staging area.** HDFS is known for linear scalability with terabytes and petabytes of data. It is also a powerful parallel processing platform that can be applied to parsing, merging, transforming, and preparing massive data sets.

- **Hadoop is also suited to data archiving.** In many data warehouse environments, the data staging area doubles as an archive of detailed source data; in many cases, the data volume of this archive exceeds that of the actual warehouse. Hadoop can be a good choice if you're archiving large amounts of raw data as many organizations do for analytics.

- **Hadoop scales with pushdown processing.** The popular ELT practice usually pushes data processing into a target relational database such as that under a DW or operational data store. However, many types of pushdown processing also work (and at massive scale) with Hadoop.

- **Hadoop can off-load your DI platform or hub.** Hadoop frees up capacity on the hub that can be applied to other DI routines or new solutions, thereby helping the hub to scale up.[2]

Hadoop aside, other relatively new data platforms can contribute to data processing in a DI context. For example, much pushdown processing is inherently relational, which Hadoop is weak with today. However, most columnar- and appliance-based platforms are relational and are optimized out of the box for such pushdowns. In another example, early adopters are using NoSQL databases for processing schema-free and unpredictably structured data (as is typical of new data sources such as sensors, Web applications, and social media).

[2] For additional details about Hadoop's uses in multiple data architectures, see the 2014 TDWI Best Practices Report *Evolving Data Warehouse Architectures* and the 2015 TDWI Best Practices Report *Hadoop for the Enterprise*, both available as free downloads at www.tdwi.org/bpreports.

☑ NUMBER FIVE

KEEP ADDING MORE RIGHT-TIME FUNCTIONS AS YOU MODERNIZE YOUR DATA INTEGRATION SOLUTIONS

Terms such as *real-time analytics*, *near-time dashboards*, and *right-time reports* are misleading. Most of the time it's not the analytics, dashboards, or reports that are real time, near time, or right time. It is usually the data integration infrastructure and its specialized interfaces that move data fast and frequently. Likewise, business methodologies such as operational BI, zero-latency enterprise, and business performance management all depend heavily on DI's real-time functions. For the sake of these popular and important technical and business practices, users in many contexts continue to modernize DI—but also their reporting, analytics, and data warehouse platforms—to infuse them with more real-time functionality.

The term *right time* assumes that there are many speeds and frequencies needed because every step in a business process (or every datum in a database) can have its own degree of urgency or freshness time frame. This is why modernizing DI for right time involves supporting several technical functionalities. These include high performance (for queries, dashboard refresh, warehouse load), microbatch (running frequently during the day to complement overnight batch processing), and data federation (to fetch small amounts of data for time-sensitive metrics). Many functions are flexible and can be configured to run at multiple right-time speeds, as with data replication, data sync, and changed data capture. If batch processing is the low end of right time, then the other end involves "true real time" (millisecond responses), as enabled by tools for event processing, complex event processing, operational intelligence, and stream processing.[3]

That's a lot of right-time functions and choices. Luckily, modern data integration platforms support multiple tool types and functionalities in one integrated development and deployment environment. Users of these integrated multi-tool environments have many options at their disposal, so their solutions can handle data at the right speed or frequency.

You probably noticed that many of the modern practices in data integration discussed earlier in this report have a right-time requirement:

- **Data ingestion** relies on many right-time rates, from traditional overnight batch processing to the continuous ingestion required of stream processing, plus many right-time gradations in between.
- **Data prep** could theoretically tap any kind of DI function, plus those for data quality, but it tends toward near-time techniques such as data federation and microbatch.
- **Data exploration** (like other variations of self-service data access) assumes an immediate response for the user, which is usually fulfilled via high-performance queries.

[3] For a more detailed discussion, read the 2014 TDWI Best Practices Report *Real-Time Data, BI, and Analytics,* available as a free download at www.tdwi.org/bpreports.

☑ NUMBER SIX

MODERNIZE YOUR DATA INTEGRATION FUNCTIONALITY TO FINALLY GET BUSINESS AND ANALYTIC VALUE FROM MULTI-STRUCTURED AND NON-TRADITIONAL DATA

We've all been paying lip service to it for years, saying we know there's valuable information in data types that are not the usual structured or relational formats. Yet few organizations have taken action, much less used these formats in production. Users interviewed by TDWI regularly talk about how they have mature skill sets and tool portfolios for relational data and a few other types of structured data, plus the interfaces associated with these. The catch is that these do not apply directly, as is, to non-traditional or "novel data"—that is, data that's outside the established structured paradigm.

However, a successful starting point is to modernize data integration skills and tools to enable functionality that's key to wringing business value from new big data and other exotic data:

Capture: Streaming data is the extreme case. Stream sources (mostly machines of different types) *push* data to your DI environment, which is backwards from the usual *pull* paradigm that DI solutions take. As a result, your DI platform needs interfaces that can capture the large numbers of small messages that most streams generate, and then store or process them appropriately. This is key to getting business value from the exploding number of sensors that are built into or are being added to almost everything on the Internet of Things (IoT), including oil wells, trucks, railcars, shipping palettes, physical plants, traffic intersections, and hand-held devices.

Other capture scenarios are more familiar to traditional DI cases. For decades, DI solutions have picked up and processed flat files containing lightly structured data, usually files containing a table dump, application log, changed data records, or a data exchange document. Today, file-based data is exploding due to the increased use of standardized file formats (e.g., XML or JSON) and logs from both enterprise and Web applications. Organizations have long acquired third-party data for consumer demographics, but many now also acquire social media data, which has its own formats. Users need modern DI platforms that can capture and natively handle old, new, and evolving file-based formats, plus allow developers to design their own support for non-standard formats.

Storage: If all the data coming into your DI environment is fully or almost relational, then storing it on a relational DBMS makes sense. However, there's already a history of failure among users who have tried to transform unique data structures to fit the relational model. Failed use cases include flattening hierarchical structures into tabular ones and storing massive amounts of human language text as binary large objects. Such practices misrepresent the original data, limit the viability of queries and searches, and obscure data lineage and audit. In a related problem, most lightly structured flat file formats transform easily and accurately into relational tables. This isn't universally good because of the overhead of transformation and the relatively high cost of relational storage.

The trend in DI-driven landing and staging is to store data in its original form when possible so the data can be processed and transformed in new ways when new application requirements arise. This way, data applies to more situations instead of being limited by storage formats that misrepresent the data and inhibit exploration and discovery analytics.

Processing: These issues are among the reasons users are deploying a wider range of data platforms, as discussed earlier in this report. The point of diverse data platforms is to address the multiple storage and in-platform processing requirements of today's multi-structured and non-traditional new data. These issues impact both data warehousing and data integration, which is why their overlapping architectures increasingly share new data platforms, from appliances to Hadoop.

Storing data in native formats on new platforms is made even more practical by new data platforms (based on columns, appliances, Hadoop, NoSQL) that can process massive data sets *in situ* with little or no preprocessing or data movement. One of the strongest modernization trends (affecting DI, analytics, DWs, etc.) is to bring algorithms and other processing logic to the data instead of the older habit of moving data to a processing tool. New platforms were built for this, and older relational brands of DBMSs have been retrofitted with in-database analytics and other *in situ* processing.

Structure: Despite the trend toward in-place processing, there are still many scenarios where an independent tool needs to access and process data on a variety of platforms. As a special case, consider tools for text mining, text analytics, and other forms of natural language processing (NLP). This class of tools is optimized for file-based data, and most tools have tight integration with Hadoop. Because the data being operated on has grammatical structure—but not relational structure—these tools are often configured and programmed to parse human language and generate data structures that can be read by other tools. These structures range from records that go into a fact table to neural net and graph structures. (Even so, other use cases may prefer a keyword search index as the tool's output.)

Because of the prominence of use cases that transform text into structured data, NLP tools are sometimes called "ETL for text," and hence are joining ETL and other DI tool types in the portfolios of modern data integration teams. The lesson here is that imposing just enough structure on unstructured data produces an output that many tools and users—both old and new—can consume for the greatest business value. Purists may scoff, but this is consistent with most analytics, which regularly reveal structures, relationships, and correlations that weren't explicit in the original data format.

Metadata: Many types of new and big data are schema-free, and their sources do not expose an accessible metadata repository or data dictionary. This is typical of many sensor feeds and anything involving dumps of human language text. Furthermore, the implicit structure may evolve unpredictably or have many variations, as seen in JSON documents. This is a challenge to traditional tools—and traditional workers!—where data access and load relies heavily on known metadata.

Even so, metadata still plays an important role with the new formats of big data and other exotic sources. However, instead of knowing and developing metadata before building a solution, metadata may be deduced *ad hoc* at runtime from implicit configurations of values found in the otherwise unstructured data. This is sometimes called "schema on read." Once a structure is discovered or deduced, a developer or an automatic tool function can capture and improve the metadata. Some metadata is relevant over time and so should be recorded in a repository, whereas other metadata only applies to one runtime session and so may be used and discarded. This is one of the most profound adjustments seen in data integration modernization—forward-facing solutions must support both traditional and new approaches to metadata management.

☑ NUMBER SEVEN

CONSIDER MODERNIZING YOUR DI TOOL PORTFOLIO WITH AN INTEGRATED PLATFORM OF MULTIPLE DATA MANAGEMENT TOOLS

A few years ago, a TDWI survey about next-generation data integration asked users if they're "using a DI tool that's part of an integrated suite of data management tools from one vendor." Only 9 percent of respondents reported using one, although 42 percent would prefer one. The same survey asked what would drive users to replace their primary DI tool. The most popular answer was: "We need a unified platform that supports DI, plus data quality, governance, MDM, etc."[4]

Since then, TDWI has interviewed many users who have abandoned a best-of-breed approach in favor of a unified toolset, making this the strongest trend among users modernizing their DI tools. This is also a strong trend among the leading DI vendors; these vendors have responded to user demand by supplying additional data management functions in a tightly integrated single platform.

Such an integrated platform typically has a strong DI and/or data quality tool at its heart, with additional tools for master data management, metadata management, stewardship, governance, changed data capture, replication, event processing, data services, data profiling, data monitoring, and so on. As you can see, the list can be quite long, amounting to an impressive arsenal of related data management tools and tool features. However, the arsenal is a mere suite—not an integrated platform—unless the tools are integrated tightly in a way that enables modern practices.

For example, as more user organizations coordinate diverse data management teams and their solutions (as in a competency center), it makes sense for the consolidated team to use a single platform for easier collaboration. Coordinated teams of this sort generally want to share meta- and master data, profiles of data sets, business rules, quality metrics, logic, and other development artifacts.

As another example, consider integrated toolsets from some vendors that enable users to design one "data flow" or similar construct, which at various steps in the flow performs functions for ETL, data federation, data quality, and master data management. Users like these modern designs because they reflect the real-world fact that most data being integrated needs multiple improvements, standardizations, merges, and deduplications. Attempting a unified data flow with a best-of-breed tool portfolio is problematic because of the challenges of making multiple tools from multiple vendors interoperate reliably with high performance and rich functionality.

[4] See the discussions about Figures 2 and 5 in the 2011 TDWI Best Practices Report *Next Generation Data Integration*, available as a free download at www.tdwi.org/bpreports.

Although comprehensive, an integrated DI platform is rarely the only tool product in use:

- Many user organizations have a primary DI tool or platform, which is the standard for most solutions, especially those of enterprise scope. They may also have secondary tools that are simpler, cheaper, and applied to smaller projects, especially departmental ones.
- With all the changes occurring in data management, a team may need additional tools for new data (e.g., NLP for text data) or new data platforms (especially Hadoop).

Whether you go best of breed, integrated DI platform, or a mix of both, demand that your tool provider stay up to date with support for new data sources and data types, as well as interfaces to new data platforms and in-platform processing for them.

CONCLUSION

Let's review the seven high-priority issues in data integration (DI) modernization discussed in this report:

Multiple data ingestion techniques allow data to move at its own speed or generation frequency. That way, data arrives in target data platforms as soon as possible and is available for immediate business use in dashboards, reports, and analytics.

Data prep enables a data analyst, data scientist, or similar user to construct a data set prototype quickly without being slowed down by excessive modeling and standardization. Such speed is critical to modern practices in analytics.

Self-service data access helps users work with spontaneity and speed because they aren't waiting for IT or a data management team to construct a data set for them. This is key to modern practices such as agile development, data exploration, and data discovery.

New data platform types, when incorporated into a modern data integration infrastructure, provide new options for capturing non-traditional data and massive volumes of data, as well as for analytic processing and DI transformations.

Right-time data movement is the secret sauce that accelerates many time-sensitive business practices, including operational BI, performance management, and a wide range of real-time analytics. Because there are many "right" times for moving data, proper enablement typically involves multiple data integration functions that operate at multiple speeds and frequencies.

Non-traditional data promises great business value for decision making and analytics. To support that goal, a modern data integration platform must capture data pushed to it, handle unstructured data types, support new approaches to metadata, and coordinate with tools for natural language processing.

Integrated tool platforms include many tool types for data integration, data quality, and master data management. The tools are tightly integrated to facilitate collaboration among developers and to foster the design of modern DI solutions that call multiple, highly diverse tool functions.

ABOUT OUR SPONSORS

informatica

www.informatica.com

Informatica is a leading independent software provider focused on delivering transformative innovation for the future of all things data. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. More than 5,800 enterprises depend on Informatica to fully leverage their information assets residing on premises, in the cloud, and on the Internet, including social networks.

sas

www.sas.com

SAS Data Management is an industry-leading solution built on a common integrated platform that helps you improve, integrate, and govern your data. No matter where your data is stored—from legacy systems to Hadoop—SAS Data Management helps organizations access the data they need.

Organizations that are modernizing legacy hardware/software systems have found SAS Data Management indispensable in integrating and managing a variety of data from both structured and unstructured sources. With core offerings such as SAS Data Loader for Hadoop, SAS Event Stream Processing, SAS Data Management, and SAS Data Federation, SAS meets the new business requirements outlined within this report.

ABOUT THE AUTHOR

Philip Russom is director of TDWI Research for data management and oversees many of TDWI's research-oriented publications, services, and events. He is a well-known figure in data warehousing and business intelligence, having published over 500 research reports, magazine articles, opinion columns, speeches, Webinars, and more. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and BI consultant and was a contributing editor with leading IT magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data warehousing solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide Membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

ABOUT TDWI CHECKLIST REPORTS

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.