research

CHECKLIST REPORT

2018

The Cloud Data Integration Primer

Get Started Integrating Your Data in the Cloud

By Philip Russom

Sponsored by

talend

tdwi

Transforming Data With Intelligence™

JULY 2018

TDWI CHECKLIST REPORT

The Cloud Data Integration Primer

Get Started Integrating Your Data in the Cloud

By Philip Russom

Transforming Data With Intelligence™

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

A number of newly mature trends are making cloud-based data integration platforms, technologies, and user best practices more relevant than ever.

- **The cloud is a well-established platform.** The cloud has become a preferred computing and data platform for modern applications, data, and data-driven business practices such as analytics. The cloud is also becoming prevalent for data management disciplines, including those for data integration and related fields such as data quality, master data management, data warehousing, and reporting.
- **Many organizations use multiple clouds.** Due to cloud's successful proliferation, many organizations are now "multicloud" because they use multiple brands of software-as-a-service (SaaS) applications (e.g., Salesforce and Marketo) in addition to traditional enterprise applications. This means they manage data on both cloud and on-premises systems. This complex system environment presents unique cloud-to-cloud and hybrid challenges and opportunities that data integration in the cloud can handle and leverage.
- **Data from the cloud and the Internet now coexists with enterprise data.** Many organizations have data sources and targets both on premises and in the cloud because they have embraced big data, social media, the Internet of Things (IoT), SaaS applications, and cloud storage. To accommodate the resulting hybrid data environment, future-facing organizations need to modernize and extend their integration infrastructures to support the Internet and the cloud fully.
- **Cloud development tends to be agile.** The speed of business continues to accelerate, demanding that data, sources, and data-driven products (e.g., reports and analyses) be delivered for business use in minimal time. Cloud-based integration platforms with agile interfaces can compress development cycles by incorporating new data sources and users quickly.
- **Self-service is a data requirement.** Businesspeople and other users are demanding self-service access to data lakes and other cloud-based data sets for analytics so they can perform self-guided data exploration, data prep, and visualization. In a related trend, the only way to scale to increasing numbers of data consumers is to enable them with self-service data access. A cloud-based data integration solution can provide a metadata-driven central point for sharing data and collaborating.

Users need to address these trends if they are to leverage them for organizational advantage by operating, competing, and innovating based on data.

The problem holding them back is change management; CIOs, CDOs, and architects know they need to move some of their data operations to the cloud but they fear the risks, cost, and danger of disrupting daily business. Users facing these changes need a primer on cloud data integration to get them started.

This report is that primer. It offers several recommendations about how to conceptualize cloud data integration, what capabilities to look for, and how to align cloud data integration efforts with business goals and requirements.

NUMBER ONE

LEARN WHAT CLOUD DATA INTEGRATION IS AND DOES

What is cloud data integration?

Cloud data integration is a secure, cloud-oriented integration platform that combines the power of traditional data integration capabilities with modern and agile approaches to data-driven solution development with a focus on extending data integration infrastructure to natively support the cloud. It also supports related functions for data quality, master data, metadata, and event processing plus handling big data, IoT data, and other new data sources or targets from the cloud or the Internet.

What does cloud data integration do?

Modern, comprehensive data integration can run anywhere—on premises or in the cloud—to liberate siloed systems and to provide the business with the greatest data value. To achieve these goals, data integration in the cloud enables developers to design unified solutions that can run natively for local performance or functionality advantages. For example, an emerging architecture for hybrid data integration controls solutions from the cloud and operates on cloud-based data but also spawns remote engines or agents on on-premises servers or hardware. Because user organizations increasingly have multiple cloud-based applications and data sets, data integration in the cloud should likewise support the native interfaces, calls, and data models of SaaS applications, cloud storage, and cloud data warehouse platforms.

What is the point of data integration in the cloud?

It addresses a real-world need for data integration infrastructure and solutions that reach all applications, data, and people regardless of their types or locations. Comprehensive data integration of this scope is mission-critical as enterprises of all sizes and industries deal with increasingly hybrid and distributed data environments.

Obviously, data integration in the cloud helps user organizations establish enterprise-grade integration solutions for cloud environments. It also helps users migrate to the cloud, including the migration of multiple applications, data sets, and user constituencies.

Data integration in the cloud can also be a modernization strategy that maintains enterprise investments in traditional approaches to data integration (because the business need for those will not go away) while extending into modern hybrid and cloud-driven approaches (which enable the business to fully leverage new information assets and resources).

NUMBER TWO

EMBRACE CLOUD DATA INTEGRATION FOR ITS COMPELLING USE CASES

TDWI regularly sees users succeeding with cloud-based data integration solutions. Their success points to a number of compelling use cases that also make low-risk starting points for an organization's initial efforts in cloud data integration.

Advanced analytics. One of the strongest drivers for change in data management today is the demand for a wider range of approaches to analytics. Businesses need more diverse analytics to support their efforts in profitability, growth, competitiveness, and customer acquisition and retention. Most firms already have mature implementations of online analytical processing (OLAP), but they need more discovery-oriented or predictive approaches such as those enabled by mining, statistics, graph, and machine learning.

Many organizations are choosing to extend their portfolio of advanced analytics by using cloud-based systems. This begins with cloud data integration to pull together large volumes of data from diverse sources as required for the cross-source correlations that most advanced analytics tools operate on. Though a mix of cloud and on-premises sources may be involved, cloud storage is increasingly the point where large analytics data sets are aggregated or persisted in a data lake. In turn, working with cloud storage requires support for many interfaces and cloud providers, as provided by modern platforms for cloud data integration.

Cloud data warehousing. The modern data warehouse is, by definition, a compilation of numerous data collections, each focused on dimensional models, data domains of interest, time series, marts, operational data stores, and so on. As users make decisions about how to modernize their warehouse, they migrate some data collections from legacy on-premises databases to cloud-based data platforms. Hence, some data warehouses today use a mix of platforms. A growing number of users choose to migrate the whole warehouse to the cloud. Whether wholly or partially on a cloud platform, the modern data warehouse needs cloud data integration to migrate warehouse data to the cloud initially as well as to feed the warehouse from hybrid sources during daily production.

Operational reports that tap Internet sources. The time-honored practice of operational reporting continues today because few managers can run their enterprises without the daily updates that reports provide. Even so, operational reports have changed dramatically in recent years, especially in organizations using SaaS, Web, and IoT applications. With so much valuable operational data originating from multiple clouds and the Internet—sometimes synchronized across clouds and on-premises systems—modern

operational reports would not be complete and up to date without cloud data integration.

Multicloud data sync. Some of the most popular SaaS apps today automate sales and marketing business processes—plus other customer-facing functions such as customer service, billing, and shipping. Achieving complete customer views in a multicloud environment—and synchronizing customer data across related customer-oriented applications (whether on premises or in the cloud)—demands sophisticated and high-performing cloud data integration. This use case—hybrid and complex in the extreme—is usually *not* a starting point for the average user organization, but it should be kept in mind as a future goal for cloud data integration programs as they mature.

NUMBER THREE

LEVERAGE CLOUD TO ENHANCE DATA INTEGRATION AND VICE VERSA

Several cloud capabilities can enhance data integration solutions in important ways.

Cloud's elastic scalability. Many data integration workloads ramp up quickly, demand considerable server resources, and then subside just as quickly. Common examples include data ingestion, data transformations, and preprocessing data prior to loading targets. When these occur, an elastic cloud can automatically marshal needed resources, then reallocate resources after intense data integration workloads complete.

Cloud centralization of semantics and collaborative capabilities. There is a long, successful tradition in data integration architectures, a tradition that centralizes such elements as metadata, other semantics, development routines, data profiles, business rules, and quality metrics. Centralizing shared resources and services makes data management consistent and governable while increasing developer productivity and collaboration. These practices originated on premises but are now available on the cloud, too, so that resources and services can be shared even more broadly among geographically dispersed people and departments, as well as applied in production among the multiple platforms of hybrid data integration workflows.

Cloud's favorable economics. Server and storage resources tend to cost less on cloud platforms compared to traditional on-premises resources. Furthermore, the cloud provider handles server capacity planning, optimization, upgrades, and maintenance, taking those time-consuming distractions off the plates of data management professionals. Finally, by using cloud-based servers and storage, data management staff need not devote time to system integration or burn up budget on capital expenditures.

Data integration's support for cloud falls into two categories:

- **Data integration running natively in the cloud.** As noted, data integration processes executed in the cloud benefit from cloud's scalability, neutrality, and affordability. Furthermore, this architecture places data integration upstream near clouds and other Internet-based sources and targets, thus enabling new practices in ingestion, triage, real-time, and streaming.
- **Data integration interoperating with multiple clouds.** To succeed with the extreme complexity of today's hybrid data environments, you need deep support for new technologies (such as Spark) as well as open source tools, big data sources, and cloud storage. For the greatest speed and scale (plus richest functionality), cloud data integration must also support interfaces to popular SaaS apps. For use cases in analytics and data warehousing, cloud data integration must also support interfaces to cloud-based data warehouses and other databases, the popular ones being AWS EMR and Redshift, Azure HDInsight and SQL Data Warehouse, Google Dataproc and BigQuery, and Snowflake.

NUMBER FOUR

KNOW WHAT TO LOOK FOR IN A CLOUD DATA INTEGRATION PLATFORM

Data is changing rapidly and radically, thereby challenging data management professionals' skills and solutions. At the same time, businesses are under renewed pressure to leverage data, analytics, and new platforms such as cloud to remain competitive, efficient, and growing. These and other factors (as discussed in the Foreword of this report) are driving many organizations to modernize and expand existing data integration infrastructure, whereas others choose to "replatform" by ripping out older tools and replacing them with a modern data integration platform. For organizations facing platform extensions or replacements, the following rundown of desirable cloud data integration capabilities can guide their choices.

Mature, comprehensive data integration platforms share these key characteristics:

- **Running in the cloud and interoperating with popular clouds.** As discussed in the previous section, data integration's deep support for clouds falls into two categories. Both are needed by modern organizations that are dealing with today's mix of traditional and modern systems. To meet

growing technology and project demands, a good cloud data integration platform must enrich job execution in the cloud to ensure smooth, continuous integration and deployment, regardless of the mix of systems on premises and in the cloud.

- **Unified toolset on a single platform.** Whether on premises, in the cloud, or a mix of both, a truly comprehensive data integration platform should support multiple data management and integration disciplines, including data integration, data quality, master data management, and event processing. Most data developers today are cross-trained and regularly work with multiple data disciplines, so a unified toolset is more productive for them and fosters consistent data standards and unified flow designs.
- **Hybrid integration workflows.** The old way of designing data integration solutions was to create a plague of small routines—each focused on one function of one data discipline—then stitch them together via scheduling. Instead of this piecemeal approach, today's modern data integration solution often consists of one unified workflow that calls a wide range of integration, quality, transformation, messaging, and interfacing functions. Dependencies among calls are easier to see and manage than in the piecemeal approach. Scheduling, monitoring, and optimization are greatly simplified, too.

  The single, unified workflow for data integration is now morphing into the hybrid integration workflow as users need to incorporate data from new sources and targets including clouds, the Web, and the IoT. For hybrid integration workflows to succeed, users need data integration in the cloud, which supports new big data, social media, and IoT data, at scale near the new sources and targets.
- **Data security, privacy, and protection.** As data integration platforms and solutions expand to address new data opportunities and hybrid environments, we must continue to satisfy fundamental requirements such as data security. Security isn't just for data at rest; data integration infrastructure must also secure data in motion, which explains why TDWI sees user organizations increasing their use of data encryption, data masking, and digital certificates as part of their data integration solutions.

  A mature cloud data integration platform must have built-in data masking, de-identification, and obfuscation capabilities. These are essential for complying with rigorous regulations and policies, especially the new General Data Protection Regulation (GDPR) from the European Union. Having these capabilities native in a cloud data integration platform means they can also apply to the management of data testing and data privacy requirements when sharing data with external parties.

Leading-edge data integration capabilities soon to be commonplace

- **Machine learning.** Today, machine learning (ML) most often supports predictive analytics applications. This is an important practice, and it will continue. However, a new trend embeds ML-driven intelligence into data integration tools. Such embedded ML algorithms and predictive models provide automation for well-understood but time-consuming development tasks such as mapping sources to targets, cataloging data, and onboarding new sources. ML can also help with the optimization of data integration system performance and capacity management.
- **Object storage.** For data-driven use cases, object storage is preferred over a block or file approach for cloud-based storage. Object storage is similar to a database in that it supports volume definitions, and data objects have metadata, so object storage feels familiar for data professionals and easily interfaces with tools for integration, reporting, and analytics. For example, a cloud data integration platform can tap cloud-based object storage for established data integration tasks, such as data staging, in-place processing, and data prep, plus emerging tasks such as modern data pipelining, serverless data management, dockerized execution, and data formatting and enhancement via Avro and Parquet.

NUMBER FIVE

ENABLE SELF-SERVICE DATA PRACTICES VIA A CLOUD DATA INTEGRATION PLATFORM

Self-service access to data is a high-priority requirement for many types of users. Therefore, data management professionals need to build support for self-service into most solutions, including those for cloud data integration. A growing number of business users (who know the basics of data structures and how to create queries) want to work with data hands-on and independently, with little or no time-consuming support from IT or data management personnel.

Data-savvy business users need to explore, prep, visualize, and analyze data to understand which business entities and processes are represented in the data before they decide how that information can be applied to improving business processes and decision making. With so many new data sources coming online nowadays—involving SaaS apps, social media, big data, and IoT—unprecedented numbers of business users are working with data in a self-service manner. Business users aside, increasing

numbers of technical users avail themselves of self-service access, as when a data analyst or data scientist needs a quick-and-dirty read of data before deciding which direction to take with a new analytics assignment.

Note that many of the characteristics that enable self-service data practices also enhance agility and productivity for both business users and technical developers. For example, self-service can enable fast-paced technical practices, such as agile or lean development. In these methods, a prototype data set is required very early in a project, and it can be created quickly and easily via self-service data access, exploration, discovery, and data prep.

Furthermore, many technical managers favor self-service because it crowdsources some data integration work, thereby offloading part of the technical team's workload. Business managers are likewise supportive of self-service because it helps businesspeople come to useful insights faster and more creatively.

Requirements for making modern self-service data practices work

- **An easy-to-use tool for end users.** Without ease of use, nontechnical users are limited in terms of how sophisticated their queries and discoveries will be. Technical users get a productivity boost from ease of use.
- **Business metadata or equivalent semantics.** This is the foundation for all data-driven self-service practices, and nontechnical users will not succeed without it. The collection of business metadata may be organized as a metadata repository, business glossary, or data catalog.
- **Interfaces to data sets of interest, whether on premises or in the cloud.** As with everything else we do in data management, data exploration and other self-service data practices will reach into many data sources, thereby requiring interfaces to those, as well as data merging or virtualization for distributed queries.
- **A data integration platform that provides most of the above.** The end-user tool described above may be a tool for BI, data visualization, or analytics. However, some data integration platforms also include easy-to-use tools. Of course, the semantics and interfaces described above usually require a mature data integration platform, and that platform should be up to date for emerging cloud and Internet data sources and targets.

NUMBER SIX

USE YOUR CLOUD DATA INTEGRATION PROJECT TO LEARN NEW PRACTICES YOU CAN APPLY ELSEWHERE

Your initial endeavors with cloud data integration may start as a primer, but it doesn't have to stop there. Lessons learned can have far-reaching impact.

For example, cloud data integration forces a data management team and their business users to do things differently, using different toolsets, technologies, and data types, typically in shorter timeframes and with less rigid adherence to dogmatic traditional standards. Working with cloud data integration (or any new data platform) can be a metaphoric proof of concept. New methods of work proven in one project or platform can be applied to others, even older solutions.

Leverage your experience with cloud data integration to break old technology bad habits, which too often take months to develop new data models and interfaces or to integrate systems. If new practices are applied to older data practices, the result can be a more agile and productive team.

Liberate businesspeople so they can get to know data and the new entities it represents, then think "outside the box" to innovate with new ways of applying information—faster, smarter, and broader—to business operations, planning, strategy, customer relations, partner programs, and other groups.

Change is good when it improves an enterprise and its use of data. Yet change is hard to kick-start and difficult to sustain. To get beyond the barriers, we must foster any project that makes change easier. Cloud data integration is one of those projects.

Embracing cloud data integration is not only about taking advantage of newer, faster, leaner, and cheaper technology. It's also about modernizing how we think and work and excel—on both technology and business levels—so the organization is better positioned for success in the near and far future.

SUMMARY

In this report we have discussed how data, its management, and its business use cases are evolving aggressively as more data sources and applications come online involving the Internet and clouds. In response, many enterprises need to extend or replace their data integration infrastructure to fully support enterprise, Internet, and cloud-based sources and targets. The goal is to provide seamless integration and interoperability across all data sets and applications, even in extremely hybrid environments. Achieving this goal requires modern cloud data integration, and this report's checklist recommendations (summarized below) provide a primer for cloud data integration.

Learn what data integration in the cloud is and does. Cloud data integration addresses a real-world need for data integration infrastructure and solutions that reach all applications, data, and people, regardless of their types or locations. This is because enterprises of all sizes and industries are dealing with increasingly hybrid and distributed data environments. Cloud data integration also helps users to migrate to the cloud, including the migration of multiple applications, data sets, and user constituencies.

Embrace cloud data integration for its compelling use cases. TDWI increasingly sees organizations with hybrid data environments using cloud data integration within strategic programs for advanced analytics, data warehouse modernization, operational reporting for SaaS applications, and data synchronization across multiple clouds.

Leverage cloud to enhance data integration—and vice versa. On the one hand, data integration benefits from the cloud's strengths, namely elastic scalability, centralization, and favorable economics. On the other hand, multicloud environments benefit from data integration's ability to synchronize, augment, and improve data across all platforms.

Know what to look for in a cloud data integration platform. The high-priority characteristics of a mature cloud data integration platform include running on all key platforms (whether on premises or in the cloud), a unified toolset for development and administration, unified (though hybrid) workflow designs, and multiple approaches to security and data protection. As cloud data integration matures further, expect to see support for machine learning and object storage.

Enable self-service data practices via a cloud data integration platform. Self-service access to data is a high-priority requirement for many types of users. Therefore, data management professionals need to build support for self-service into most solutions, including those for cloud data integration. Making modern self-service data practices work requires an easy-to-use tool for end-users, business metadata (or equivalent semantics), and interfaces to data sets of interest (whether on premises or in the cloud).

Use your cloud data integration project to learn new practices you can apply elsewhere. For example, cloud data integration forces a data management team and their business users to do things differently, with different toolsets, technologies, and data types, typically in shorter time frames and with less rigid adherence to dogmatic traditional standards. Leverage your experience with cloud data integration to break old bad habits and to make your team more agile and productive.

ABOUT OUR SPONSOR

Talend, a leader in cloud integration solutions, liberates data from legacy infrastructure and puts more of the right data to work for your business, faster. Talend Cloud delivers a single platform for data integration across public and private cloud, as well as on-premises environments, and enables greater collaboration between IT and business teams. Combined with an open, native, and extensible architecture for rapidly embracing market innovations, with Talend you can cost-effectively meet the demands of ever-increasing data volumes, users, and use cases.

Over 1,500 global enterprise customers have chosen Talend to put their data to work. More than a third of the *Fortune* 100 U.S. companies are Talend customers including GE, HP Inc., and Domino's. Talend has been recognized as a leader in its field by leading analyst firms and industry publications including *Forbes*, *InfoWorld*, and *SD Times*. Talend is Nasdaq listed (TLND) and based in Redwood City, California.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, 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 BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management 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 management 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 THE AUTHOR

Philip Russom, Ph.D., is senior director of TDWI Research for data management and is a well-known figure in data warehousing, integration, and quality, having published over 550 research reports, magazine articles, opinion columns, and speeches over a 20-year period. Before joining TDWI in 2005, Russom was an industry analyst covering data management at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and consultant, was a contributing editor with leading IT magazines, and a product manager at database vendors. His Ph.D. is from Yale. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.