research

CHECKLIST REPORT 2020

Five Steps for Accelerating Data Readiness

Developing a technology strategy to drive better analytics and governance across the enterprise

By David Stodder

Sponsored by: boomi

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

OCTOBER 2020

TDWI CHECKLIST REPORT

Five Steps for Accelerating Data Readiness

Developing a technology strategy to drive better analytics and governance across the enterprise

By David Stodder

TABLE OF CONTENTS

2 FOREWORD

4 NUMBER ONE
Advance data readiness by developing knowledge graphs of data relationships

6 NUMBER TWO
Establish a data catalog to improve location, analysis, and governance of data

8 NUMBER THREE
Address weaknesses in governance that impact readiness and reduce data trust

10 NUMBER FOUR
Improve readiness through efficient, flexible, and integrated data preparation

11 NUMBER FIVE
Modernize pipelines for analytics and AI/ML with catalogs and workflow automation

12 A FINAL WORD

13 ABOUT OUR SPONSOR

13 ABOUT TDWI CHECKLIST REPORTS

13 ABOUT THE AUTHOR

13 ABOUT TDWI RESEARCH

tdwi

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

FOREWORD

Data is everywhere, generated by an array of applications, digitally transformed business processes, mobile devices, and Internet of Things (IoT) sensors. Unfortunately, even as organizations ingest this data into enormous data lakes and cloud storage systems, much of it remains unknown.

Only a fraction of it moves through the long data quality and transformation processes that structure it into "known" data for target systems such as data warehouses, where users are able to access it for business intelligence (BI) reports, dashboards, and managed analytics. Too much remains untapped; many organizations are unable to prepare and process data in time to handle analytics workloads for solving business questions, improving operational processes, and building more valuable customer relationships. Instead, it sits in data "swamps" and becomes an expensive management headache.

This TDWI Checklist discusses how organizations can modernize technology strategies to create a faster path to readiness with all their data, not just the portion currently defined, structured, and formatted for BI reporting and analysis.

Known data is generated by traditional business applications primarily for financial and performance management. Through experience, organizations understand how to locate, access, and govern this data.

To be sure, structured data can be unknown; it can come from diverse applications and platforms. It is growing in volume and is often spread across disparate data silos, making it complicated to access and prepare. Data quality and consistency issues can render structured data almost as "unknown" as semi- and unstructured data. Users struggle to gain complete and trusted views of it, and organizations struggle to govern it.

Growing semi- and unstructured data, however, presents new challenges. Semistructured data could be JavaScript Object Notation (JSON) data used in web applications and forms and require specialized access.

Unstructured data has a format that is not known or a format the data management system is not set up to search, query, and access. Common examples include sensor data, video, image files, social media posts, and log files. Many organizations would like to use this data to provide fuller, more contextual understanding of structured data such as that generated by transactions.

When semi- or unstructured data is collected from non-transactional customer activity in different channels, it often falls into the category of "behavioral" data. If its relationship to structured transaction data could be mapped and visualized, it would be highly valuable in understanding customer preferences, buying patterns, and how customers influence each other.

Data relationships between collections of structured, semistructured, and unstructured data could also be valuable for many other types of use cases including fraud and abuse detection, equipment maintenance, situation awareness, manufacturing process control, and operations improvement.

To be ready for analytics and actionable for decisions, organizations need to make all their data easier to discover and prepare. In addition, to adhere to data privacy and other regulations,

FOREWORD CONTINUED

structured, semistructured, and unstructured data that organizations are storing and managing must be appropriately protected and governed.

This TDWI Checklist will discuss how data catalogs, knowledge graphs, graph databases, master data management, and workflow automation can enable organizations to move faster to prepare all their data, develop pipelines for analytics, and increase the data's business value.

1 ADVANCE DATA READINESS BY DEVELOPING KNOWLEDGE GRAPHS OF DATA RELATIONSHIPS

Data readiness needs to be about more than giving users access to quantities of data. Decision makers are often frustrated by standard "data dump" reports that do little to increase their knowledge of customer behavior, potential fraud and governance risks, how to optimize manufacturing supply chains, or what factors to consider in determining pricing. They need insights into relationships between different data entities and events in context. Too often, decision makers are left depending (and waiting) on developers to write complex programs to unearth data relationships. These programs tend to be brittle and hard to change when users have unanticipated questions.

Also driving demand for faster insights into data relationships is growth in artificial intelligence (AI) for analytics and the embedding of AI-driven automation in systems such as chatbots for customer engagement. AI-infused systems would benefit from data management that facilitates easier mapping of relationships between "implicit" behavioral data such as that generated by customer activity and "explicit" data that results from intentional actions such as a customer purchase or feedback using a rating system that registers stars or thumbs up or down.

Because there may be no explicit action resulting from behavior such as looking at a selection of products, reading other customers' comments, or even filling (but then abandoning) a shopping cart, the person's intentions are unclear.

AI and analytics can learn partially about customer preferences from purchases and rating systems; however, if underlying data management can help unearth relationships between explicit data and implicit behavioral data, the knowledge base for AI and analytics to work with would be more complete. AI-based recommendations could be more personalized and timely.

Most organizations rely on traditional relational data management to support BI and data warehousing systems as well as applications such as customer relationship management (CRM). These systems excel at recording explicit data and are geared to managing data that conforms to known definitions, formats, and structures. Data relationships in these systems are simple and smaller in number.

To store and manage large amounts of semi- and unstructured data as well as structured data from less familiar data sources and applications, many organizations set up data lakes using NoSQL. However, the petabytes of implicit and explicit data can quickly become vast and unknown. Simply stored in the data lake, the data is far from ready for fast visualization or analysis of complex relationships.

KNOWLEDGE GRAPHS AND GRAPH DATABASES

To search for and analyze data relationships more effectively, organizations need models and systems that can store data relationships and make them discoverable. Knowledge graphs, network-based representations, and graph databases specialize in these capabilities, enabling business users, analysts, and AI applications to navigate relationships found in implicit and explicit data. They enable organizations to examine the significance of relationships between multiple data elements.

The use of knowledge graphs, graph models, and network-based representations is not new; Google and other search engine providers use knowledge graphs to increase the accuracy, speed, and completeness of search results.

1 ADVANCE DATA READINESS BY DEVELOPING KNOWLEDGE GRAPHS OF DATA RELATIONSHIPS CONTINUED

Knowledge graphs can capture the semantic richness of data relationships beyond what traditional data catalogs provide with metadata. The richness comes from building network-based representations that are not limited to structured data; the graphs can include both explicit data and implicit data that is semi- or unstructured.

The relationships are visibly represented in graph models by "edges," the lines or arcs that connect any two "nodes" (also called "vertices") representing data objects. The edges establish semantically relevant connections—potentially many of them—between the nodes. Graph databases can then manage and retrieve these data relationships.

A graph database, a type of NoSQL database, is a good alternative for complex data and exploring data relationships. Graph databases do not limit users to only what is predefined in a data warehouse or CRM data model. Using a graph database, users do not have to modify graph data models to fit a relational database's normalized table structure. Graph databases enable developers to avoid having to write complex SQL JOIN statements to discover associations in relational databases or program special routines to convert graph structures into relational structures. Graph models are also more flexible than traditional hierarchical parent-child models.

Many graph databases support open data standards such as World Wide Web Consortium (W3C)'s Web Ontology Language (OWL), Resource Description Framework (RDF), and RDF Query Language (SPARQL), as well as ontologies that formalize definitions of industry-specific entities and related rules for data interchange.

Adherence to standards allows users to run semantic queries against any graph database to either retrieve specific data relationship information or seek answers to exploratory questions.

Organizations should evaluate graph databases, knowledge graphs, and graph models for how they can enable faster discovery and analysis of data relationships. Organizations should also examine how knowledge graphs and graph databases could expand the breadth and power of data catalogs by making it easier to locate implicit as well as explicit data and bring knowledge of data relationships into the catalog. Data catalogs are discussed in the next section.

2 ESTABLISH A DATA CATALOG TO IMPROVE LOCATION, ANALYSIS, AND GOVERNANCE OF DATA

Organizations struggle to achieve satisfactory data readiness if users have difficulty locating relevant data. Users need confidence in the consistency of how data is defined. Analysis is faster and more complete if users can easily learn how different data elements are related within the context of a topic such as financial performance or marketing campaign effectiveness.

Administrators need a resource for locating and inventorying data so they can improve data quality, track data lineage, and fix data errors and irregularities. Administrators also need the ability to monitor all sensitive data use to reduce the risks of regulatory and security exposures.

These requirements are driving interest in establishing and modernizing data catalogs. A data catalog provides a centralized and trusted collection of metadata (i.e., data about the data) gathered from different data sources to enable users, developers, and automated applications to learn where data is located and how it is defined and structured.

About half of organizations surveyed recently by TDWI (51 percent) say that establishing a data catalog is one of the most important steps they could take to improve users' success with BI and analytics.[1] Three-quarters of these organizations (75 percent) want a data catalog to make it easier for users to search for and find data; over half want to use it to improve data governance, security, and regulatory adherence (57 percent).

Not all data catalogs are enterprise-level catalogs run by IT. Individual departments such as marketing, sales, or finance might establish a data catalog to centralize metadata and other information about data sets specific to their needs. For example, a data catalog could help insurance sales managers and representatives see what data elements are relevant to certain policies, making it easier to gain a complete view of all the data and understand how data elements are organized within topic hierarchies. Some data catalog systems can centralize metadata information from semi- and unstructured data to make it easier to find and relate documents using XML or JSON formats or log files in a data lake.

As a second option, some organizations choose to establish a data catalog to support a specific analytics project so team members can move faster to locate data in multiple systems and resolve discrepancies in data definitions. Team members could use a feature in some modern data catalogs that enables users to add their "tribal wisdom" about different data sets through comments and annotations; this crowdsourced information can be valuable to project team members, alerting them to data quality or consistency issues. Teams can also use crowdsourcing functions to curate data for other stakeholders.

A third option some organizations choose is to set up a data catalog and apply crowdsourcing functionality for all their data stored and managed on a specific cloud platform such as Amazon AWS, Microsoft Azure, or Google Cloud Platform.

Although department-, project-, platform-, and data source-specific catalogs are valuable, an enterprise data catalog that transcends disparate silos can help organizations address data definition conflicts and quality problems that impact users

[1] All research references in this report are from the 2020 *TDWI Best Practices Report: Evolving from Traditional Business Intelligence to Modern Business Analytics*, online at tdwi.org/bpreports.

2 ESTABLISH A DATA CATALOG TO IMPROVE LOCATION, ANALYSIS, AND GOVERNANCE OF DATA CONTINUED

across the organization. As noted earlier, TDWI research finds that governance and regulatory adherence are key drivers in establishing a data catalog. An enterprise-level data catalog can aid governance by documenting and tracking data lineage information about the data's origin, who was responsible for its sourcing, and how it has been collected, transformed, copied, replicated, and shared. Governance audits and regulatory reporting typically need this information.

Only 23 percent of organizations surveyed by TDWI are satisfied with how well users can employ a data catalog to find data, understand data relationships, and tap it as a knowledge asset about data and its lineage. This indicates that whether organizations have departmental, project, platform, or enterprise data catalogs, there is room for improvement. Here are three recommendations:

- **REDUCE MANUAL WORK THROUGH AUTOMATION.** Many organizations are held back in their development and maintenance of data catalogs by the amount of manual work traditionally involved. Administrators have trouble keeping up with new and voluminous data sources as well as proliferation of data silos. However, AI techniques such as machine learning (ML) plus software automation can relieve organizations of manual work in populating data catalogs, surfacing anomalies and discrepancies, and keeping catalogs up to date. AI can also drive user recommendations for data sets based on relevance, trusted quality, and department- or project-specific contexts, enabling organizations to curate data rather than just provide access.
- **IMPROVE EASE OF USE.** To make it easier for users to search for and find data, organizations should upgrade data catalogs with more advanced search and natural language processing (NLP) capabilities. AI and NLP can help users discover how data elements are related within their subject matter context. Search that uses NLP capabilities helps users refine exploration so they are locating only relevant data and are not swamped by unneeded results. Ease of use, smarter search, and crowdsourced annotation can give users a faster path to relevant data.
- **ENABLE MORE RAPID DISCOVERY OF DATA RELATIONSHIPS.** Data catalogs that make use of knowledge graphs and graph databases can reduce the complexity involved in discovering data relationships. The combination can also speed inventorying larger volumes of unfamiliar semi- and unstructured data. Along with helping users analyze data relationships, integrating graph functionality into data catalogs enables organizations to govern data more effectively. Organizations can use graph models to visualize whether analysis and sharing of particular data relationships might violate regulations regarding the privacy of personally identifiable information (PII).

3 ADDRESS WEAKNESSES IN GOVERNANCE THAT IMPACT READINESS AND REDUCE DATA TRUST

Data governance is vital to data readiness. The primary focus of data governance today is to define rules and policies for protecting and securing sensitive data such as PII as it is defined by common data privacy regulations. Governance practices and technologies must monitor data use to make sure rules are followed when data is collected, moved, copied, analyzed, and shared.

TDWI research finds that many organizations lack confidence in their ability to govern and secure all their data, especially as they move data into the cloud. Nearly half of organizations surveyed by TDWI (45 percent) cite protection of sensitive data as one of their biggest challenges when augmenting or replacing existing on-premises systems with cloud-based data platforms and services.

More than half of organizations surveyed (52 percent) regard growth in self-service BI and analytics as one of their biggest governance challenges. Organizations will often restrict data access for self-service BI and analytics if they are not confident in how well they can govern data use and sharing, an act that can frustrate users.

For users, as important as protecting sensitive data is the second mission of governance: to improve trust in the data. Our research shows the importance of this objective: 43 percent of organizations surveyed are using governance to increase users' confidence in the data and 45 percent are training users in governance responsibilities.

Central to trust is data quality, which is why governance and data quality should be closely aligned. Data quality processes include profiling, validation, cleansing, and addressing consistency and redundancy problems. Nearly two-thirds of organizations surveyed (63 percent) say they are monitoring data quality to improve trust in the data. Organizations need to set up metrics to evaluate quality over time and compare the quality of different data sources so administrators know where to focus remediation.

Expansion in self-service BI and analytics plus growth and distribution of data put pressure on organizations to do more than just acquire technology to improve data quality.

Organizations should designate data stewards from both business (to contribute subject matter expertise) and IT. Data stewards need to know the organization's governance rules and policies as well as the business contexts within which they must be applied. Data stewards can be helpful in guiding users of self-service technologies to apply governance and data quality standards. Stewards can mentor users in following governance policies as they work with, share, and use data in analytics models and visualizations.

ROLE OF DATA CATALOGS AND MDM

Modern data catalogs, discussed in the previous section, can play a critical role in improving governance and trust. Catalogs can document data lineage so that organizations know the sources of the data, what has happened to it during its life cycle within the organization, and who has been responsible at each step. An enterprise data catalog can help organizations assemble an accurate inventory of data for governance reporting and auditing even if it is located physically on distributed on-premises and cloud-based platforms.

3 ADDRESS WEAKNESSES IN GOVERNANCE THAT IMPACT READINESS AND REDUCE DATA TRUST CONTINUED

Master data management (MDM) processes and technologies are also helpful for governance and trust. MDM processes include developing information integration models focused on higher-level entities such as customers and products rather than the lower-level metadata and data definitions provided by data catalogs.

MDM helps organizations discover and document how data about entities of interest in one system relates to data in other systems. MDM enables establishment of a "golden record" of high-quality trusted and governed reference data. MDM models also make it easier for administrators to track data quality problems and reduce inconsistencies and redundancy.

Organizations can use data catalogs and MDM to improve process mapping, which is helpful for visualizing how data is created, is sourced, and flows through an organization's often complex business and application processes. The mapping can reveal where there may be risks of regulatory compliance exposures and where governance monitoring needs to be strengthened.

By enabling faster recognition of data relationships, knowledge graphs and graph databases can make it easier to create process maps for governance and data quality. Process mapping is also valuable for workflow automation, which is discussed later in this report.

4 IMPROVE READINESS THROUGH EFFICIENT, FLEXIBLE, AND INTEGRATED DATA PREPARATION

To accelerate data readiness for more types of users and use cases, organizations need to address problems in data preparation. Data preparation comprises the sequence of processes that take data from its original source through profiling, transformation, integration, cleansing, and enrichment until it is in a ready state. Although they are dependent on each other, data preparation processes in practice are frequently disconnected, creating bottlenecks causing inefficiency and delay. To address problems, organizations need technologies and practices that knit together the whole sequence while not restricting user flexibility.

Many organizations want data catalogs to play an important role in improving cohesion between different stages of data preparation; 40 percent of those surveyed by TDWI want to improve data preparation by setting up a centralized data catalog. This resource can address users' difficulties in locating all data appropriate for their projects and resolving inconsistencies. Data catalogs provide a shared resource that enables organizations to overcome problems that arise when users prepare data separately with personal spreadsheets and desktop databases. These can contain inconsistent data definitions and produce data sets that include quality errors and redundancy.

Modern data catalogs can apply AI and automation to find data faster, surface anomalies, and resolve inconsistencies—especially as data volumes become larger and include semi- and unstructured data. Organizations should evaluate data catalog systems that provide user-friendly search and NLP capabilities so nontechnical users can tap metadata and discover data more easily during data preparation processes.

With self-service BI and analytics continuing to expand, organizations need to balance users' pursuit of flexibility and self-reliance with enterprise needs for governance, efficiency, and ensuring that compute-intensive preparation processes such as data transformation have appropriate resources. Enterprise IT can use data catalogs' lineage information to improve governance monitoring of data preparation processes. Enterprise IT can also analyze which data sources and sets are used most often and apply that information to resource allocation.

The combination of data catalogs plus MDM information integration models and process planning can help organizations gain a broader, end-to-end view of users' data preparation processes and how they fit—or should fit—together. For example, organizations could integrate data preparation with MDM to enable faster access to golden record data drawn from on-premises and cloud-based systems.

Finally, organizations implementing DataOps frameworks to improve stakeholder collaboration for large-scale data pipeline development should make sure teams are making full use of data catalogs and MDM. Data pipelines often include data preparation processes for profiling, blending, cleansing, transformation, and enrichment. Data catalogs and MDM can improve visibility into where users' preparation processes are running into problems locating, accessing, and integrating data. The visibility is also important to data governance monitoring in data pipelines.

5 MODERNIZE PIPELINES FOR ANALYTICS AND AI/ML WITH CATALOGS AND WORKFLOW AUTOMATION

Business-critical analytics and AI/ML workloads put stress on organizations to scale up and manage numerous data pipelines for provisioning workloads with ready data. Data pipelines ingest data, often through streaming, from sources to target locations such as a data lake, data warehouse, or BI/analytics platform.

Some data pipelines simply stream raw, unstructured data into a data lake; others involve complex data preparation workflows that include data cleansing, transformation, and enrichment before data sets can be operationalized. Organizations also need to apply governance rules and policies to data pipelines.

Thus, organizations that may have started out with a "Wild West" of unmanaged data pipelines developed as needed by data scientists and engineers for specific projects are finding that they need a systematic approach to support workload growth and govern data properly. Among organizations surveyed by TDWI research, 24 percent say they need a major upgrade to their data pipeline development and operationalization and 46 percent say they need at least some improvement.

Data catalogs can improve how rapidly and comprehensively data pipelines locate data. Catalogs can also guide pipeline developers to use trusted and governed data where possible and to tap crowdsourced knowledge about less-well-known data sets. Organizations should evaluate how they can use modern data catalogs augmented with AI/ML and automation to improve the consistency and reusability of data pipelines for larger volumes and varieties of data. AI-driven automation enables some data catalog solutions to identify data quality issues across multiple workloads so they can either be solved automatically or brought to pipeline developers' attention sooner.

Workflow automation can help organizations orchestrate not only a higher number of workloads but also complex and interdependent processes for data cleansing, transformation, and enrichment within data pipelines. Workflow automation technologies enable organizations to develop data pipelines for provisioning numerous analytics and AI/ML workloads in an organized and repeatable way.

Administrators can use workflow automation to find and fix errors more rapidly in pipeline preparation processes. Workflow automation will be increasingly critical as decision makers use AI-driven recommendations in BI systems and applications. Recommendation functions depend on quality data pipelines, raising the risk to organizations if there are delays and errors.

If organizations are implementing MDM process mapping and DataOps frameworks, they should integrate workflow automation with these initiatives. Workflow automation can help organizations implement DataOps more effectively to integrate disparate pipeline development and reduce delays, increase reuse, and promote continuous improvement cycles. Organizations can use the combination of MDM, DataOps, and workflow automation to improve stakeholder collaboration, which is key to directing the orchestration of workflows to achieve priority objectives.

A FINAL WORD

Accelerating data readiness is essential when users are counting on quality, trusted, and well-governed data for visualizations and analytics—and AI/ML developers and data scientists count on data for new applications, services, and business insights. Yet as data becomes more diverse and voluminous, organizations have to push beyond traditional and limited data preparation processes if they are to realize the value of all their data, known and unknown.

This TDWI Checklist has discussed five steps for accelerating data readiness. These include taking advantage of newer approaches such as knowledge graphs and graph databases for capturing data relationships and establishing more modern shared data catalogs and MDM. Workflow automation also plays an important role by enabling greater speed and efficiency as data volumes rise and data preparation and pipeline processes become complex.

New technologies and practices will help organizations stay on top of data governance and quality challenges as they seek to turn unknown data sources into vital resources for innovation.

ABOUT OUR SPONSOR

boomi

Boomi, a Dell Technologies business, instantly connects everyone to everything with our cloud-native, unified, open, and intelligent platform. Boomi's integration platform-as-a-service (iPaaS) is trusted by more than 12,000 customers globally for its speed, ease-of-use, and lower total cost of ownership. As a pioneer in fueling intelligent use of data, Boomi's vision is to make it quick and easy for customers and partners to discover, manage, and orchestrate data while you connect applications, processes, and people for better, faster outcomes.

For more information, visit http://www.boomi.com.

ABOUT THE AUTHOR

David Stodder is senior director of TDWI Research for business intelligence. He focuses on providing research-based insights and best practices for organizations implementing BI, analytics, data discovery, data visualization, performance management, and related technologies and methods and has been a thought leader in the field for over two decades. Previously, he headed up his own independent firm and served as vice president and research director with Ventana Research. He was the founding chief editor of *Intelligent Enterprise* where he also served as editorial director for nine years.

You can reach him by email (dstodder@tdwi.org), on Twitter (@dbstodder), and on LinkedIn (linkedin.com/in/davidstodder)

ABOUT TDWI RESEARCH

TDWI Research provides industry-leading research and advice for data and analytics professionals worldwide. TDWI Research focuses on modern data management, analytics, and data science approaches and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of business and technical challenges surrounding the deployment and use of data and analytics. TDWI Research offers in-depth research reports, commentary, assessments, inquiry services, and topical conferences as well as 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, 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.