research

CHECKLIST REPORT

2016

Governing Big Data and Hadoop

By Philip Russom

Sponsored by:

talend

tdwi

Advancing all things data.

OCTOBER 2016

TDWI CHECKLIST REPORT

Governing Big Data and Hadoop

By Philip Russom

TABLE OF CONTENTS

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

FOREWORD

Big data presents significant business opportunities when leveraged properly, yet it also carries significant business and technology risks when it is poorly governed or managed.

For example, big data from websites, call-center applications, smartphone apps, and social media can reveal how your customers behave in diverse situations, thereby enabling modern multichannel marketing. Big data can provide larger data samples that expand existing analytics for risk, fraud, and customer-base segmentation. In an increasingly data-driven business world, big data takes operational analytics and a 360-degree view of customers to a new level. Because big data usually comes from new sources, TDWI Research often refers to it as *new data* or *new big data*. The great promise and relevance of new big data is that it can be leveraged in innovative ways to develop new insights, which in turn contribute to organizational growth, competitiveness, and operational excellence.

One of the challenges to achieving these valuable goals is that traditional platforms and tools are not designed for new big data. Without modern tools, technical employees run the risk of failure in key areas such as scalability, data structure diversity, low latency, self-service data access, security, metadata management, and enterprise data standards.

Another challenge is to capture and use big data to create business value but within the guidelines of external regulations and internal policies for data usage, privacy, and security. When these guidelines are not honored and followed, a business runs the risk of compliance violations that can lead to legal issues, fines, customer dissatisfaction, and poor brand loyalty.

Successful user organizations manage the technology and business risks of new big data by depending on the best practices of data governance (DG) and on functionality within modern data management tools that support these practices. A mature and comprehensive DG program serves and balances two general goals: business compliance and technical standards. Both have serious relevance to the governance of new big data, as described below.

Business Compliance for Regulations, Data Privacy, and Security

This goal is about the control of data and its use, with a focus on reducing liability and risk relative to data management. Organizations with an existing DG program should be able to map existing governance policies to new data; however, in some cases, new policies or revisions of legacy guidelines may be in order. For example, when a new customer channel opens and starts generating data, older policies about customer privacy may or may not apply. Ideally, this determination and any ensuing updates to DG policies should be in place before data from the new channel is captured and used.

Compliance policies created by a DG program must support business goals, such as:

- **Reducing risks by creating policies** that control the access, use, and movement of data deemed sensitive by external regulations (e.g., HIPAA and Basel II) or internal measures (e.g., the corporate privacy policy)
- **Certifying new data** to assure security, privacy, and compliance before a new application, source, or data platform (such as Hadoop) is put into production

Technical Standards for Data and Data Management Solutions

This goal involves the communal creation of enterprisewide standards for data models, exchange formats, metadata, semantics, data quality metrics, and data-driven development processes. This goal has long been about making diverse enterprise data sets more standardized and high quality so they are more easily shared across business units. With new big data, that goal still applies and helps quickly assimilate new data into the broader enterprise.

Data standards from a governance committee can support technology goals, such as:

- Extending your existing customer views with additional insights drawn from new data
- Enlarging the data samples of existing analytics applications for fraud, risk, and customer segmentation
- Enabling analytics applications that are new to you, the hot ones today being the 360-degree customer view; logistics optimization; and industry-specific cases such as patient outcomes in healthcare, predictive maintenance in manufacturing, or precision farming in agriculture

This report will examine six of the most pressing issues in governing various types of new big data. This includes both old and new platforms (including Hadoop) and both operational and analytics use cases. The goal of this report is to accelerate your understanding of new and big data, plus DG's important role in putting that data into the hands of more employees without putting enterprise information at risk or breaking with compliance.

DEFINITIONS

Big data has become a commonplace term for various new data sources and growing volumes of data at use or being created today by Web applications, sensors and other machinery, smartphones and other hand-held devices, social media, and the like. It usually refers to data sets that are so large and/or complex that traditional data processing solutions are inadequate. As a result, one of the biggest challenges for data management teams is how to cost-effectively scale technology infrastructure to keep pace with the constantly growing volume of data and the variety of data sources and formats, while effectively implementing new ways to both operationalize and extract insights from big data. Many find that success with managing big data comes from using new data platforms and processing tools that are specifically designed for it.

Data governance (DG) generally refers to a set of processes that ensure important data assets are formally and consistently managed throughout the enterprise so that data can be trusted by the people using it. DG also ensures that people can be held accountable for adverse events that happen because of low data quality. DG is about putting the right people, processes, and guidelines in place to prevent and fix issues with data so that the enterprise as a whole can become more data driven while complying with external regulations and internal policies.

NUMBER ONE

GIVE PEOPLE SELF-SERVICE ACCESS TO BIG DATA STORES, BUT WITH DATA GOVERNANCE

With so many new sources and types of big data arriving each day, many business and technical users want to get their hands on it because they feel new big data has great prospects for improving both operations and analytics. Hence, if you're an IT or data management professional, you're probably under pressure to give your "internal customers" what they want and need most—namely, self-service access to big data and other data assets.

For years we've lived with the old paradigm of relying on a few data providers (highly skilled data specialists) taking weeks or months to create data sets that other technical professionals or line-of-business managers can leverage. In recent years, this has given way to a new model, wherein a wide range of user types—from highly technical data scientists and analysts to mildly technical business analysts and managers—have the right tools and skills to do their own data provisioning autonomously. The result is greater speed, agility, exploration, and discovery than ever before for a broader set of employees.

However, for self-service data access to be truly useful for multiple employees—while allowing IT to maintain DG—those employees need tools with special functionality, such as:

- High ease of use for creating, running, and iterating ad hoc queries.
- Quick responses, even with complex queries using large data sets.
- Numerous features for metadata management (as explained in "Number Three" of this report).
- Data prep as a simple and expedient complement to more feature-rich data integration approaches.
- Ad hoc and a posteriori data modeling (or data prep) while the user explores and iterates queries, as opposed to the traditional a priori modeling typically performed in data warehousing. This way, exploration leads directly to the development of a data set, which is the pragmatic output of self-service data access.
- Seamless path to using data sets in tools for reporting, analytics, and visualization.

For self-service data access and analytics to be governable, it needs:

- Security (user identification and verification) and in some cases data protection (encryption and/or masking)
- Auditing, with records of access and usage kept (perhaps via operational metadata)
- Analytical sandboxing, which isolates sensitive data in a closed environment, thereby controlling the distribution of data sets and analyses

NUMBER TWO

INTEGRATE NEW DATA TO ENABLE BROAD BUT GOVERNED DATA EXPLORATION AND DISCOVERY

Getting started with new data differs from our approaches to traditional data, because new big data tends toward massive volumes and complexity. Technical personnel are challenged to find just the right platform for capturing and storing new data, plus processing it later for analytical and operational business use. At the same time, stakeholders want to preserve their existing

investments in data warehousing (DW) and business intelligence (BI) platforms, although these traditional platforms are not the best fit for many forms of new data and the kind of discovery analytics businesses need to perform.

Successful organizations are addressing this dilemma by incorporating new platforms into their software portfolios for DW and BI. This is why TDWI sees relational warehouses being complemented by standalone platforms based on data appliances, columnar database management systems (DBMSs), and Hadoop. With multiple data platforms to choose from, data management professionals now have the flexibility to find the right platform they need for storing and processing new big data.[1] Once new data is integrated on an appropriate platform, employees with the right tools can explore and "exploit it," if you will, in order to make new business discoveries.

Data exploration is a growing best practice because it can lead to the discovery of previously unknown facts about operations, customers, financials, or logistics. These discoveries, in turn, can fuel new analytical insights and operational improvements. Data exploration is also growing because data from new sources absolutely must be studied before it is used in downstream databases and applications. The study should result in two assessments:

1. **The technical state of new big data.** This includes the data's format, model, or schema; metadata (whether embedded or deducible at runtime); condition (quality, standardization, completeness); and details of delivery (file-based, extractable from an app, streaming, etc.). This knowledge is indispensable to the technical standards side of DG as well as the ensuing development of data management solutions.
2. **The business value and compliance of new big data.** This includes business entities encompassed in the data (partners, suppliers, customers, other parties); implications for regulatory compliance and data privacy; and ways in which the business can use the data to create organizational advantages. This knowledge is indispensable to the business compliance side of DG as well as to getting full business value and ROI out of the new data asset.

Consider that these two assessments may require different tools. Studying the technical state of data may require hefty data profiling functions typically found in leading tools for data integration and data quality. However, the study of the business value and potential use of new data may be done with the self-service data exploration capabilities built into a wide variety of tools for data integration, analytics, and the latest generation of data visualization. Note that data exploration may also feed directly into related functions, especially those associated with self-service such as data prep, data visualization, and some forms of analytics. Across the board, these diverse tools and individual functions should be secure and auditable for the sake of DG.

☑ NUMBER THREE

DEVELOP METADATA FOR BIG DATA, FOR THE FULLEST USE AND GOVERNANCE

Any data professional can tell you that we have traditionally depended on metadata to describe data and facilitate data access. We say "metadata" as if it's a monolithic thing. Still, in truth, our dependence on it requires having multiple forms, including technical metadata (for automated software access), business metadata (human language descriptions that business people can understand), and operational metadata (which records details about a data access event).

Note that metadata is a critical tool for DG because it's hard to govern what you can't name and describe in a standard and sharable fashion. Furthermore, some DG approaches keep an inventory of governable data that's recorded in a shared metadata repository; in that regard, metadata management makes governing data simpler, more automated, and more accurate. Finally, operational metadata is becoming instrumental to recording data lineage and tracking data access by unauthorized employees and applications.

TDWI Research believes that the dependence on metadata will continue to grow with the advent of new data despite the fact that some sources of new data don't have obvious metadata. For example, data from streams or from Internet sources rarely has a source system that will let you extract metadata from it. JSON and XML files may (or may not) have a header from which metadata can be deduced by some software tools. A lot of new data comes to you as log files, whether created by a Web server or captured from a stream and appended to a log file. In this scenario, there's no metadata per se, but the simple record structures of logs are easily transformed into metadata. For these reasons and others, you must develop metadata for new big data.

However, for users to survive the volume, diversity, and continuous evolution of new big data, there's a pressing need for smarter tools that can be prescriptive rather than declarative. In other words, tools need to automatically infer metadata from the data itself. The tool could lead a user through a guided process, where a user accepts or revises the tool's suggested metadata (perhaps during a data exploration session). In other scenarios, the tool could proceed without human intervention based on business rules for data definitions (as when an unknown message structure appears in a data stream).

This is sometimes called "schema on read." Note that this form of metadata automation needs to work during exploration, profiling,

[1] The strong trend toward multiplatform data warehouse environments is discussed in the TDWI Best Practices Reports *Evolving Data Warehouse Architectures* (2014) and *Data Warehouse Modernization* (2016), online at www.tdwi.org/bpreports.

and development as well as in production at runtime. This report earlier mentioned data modeling on the fly in the context of self-service data access and ad hoc data-set creation. Here we see that metadata is likewise developed and managed on the fly because metadata is required to define and share the created data set and its model.

There are other reasons why developing and managing metadata for new big data is important:

- The repurposing of new data regularly takes data sets to people and applications that require metadata-driven tools and interfaces.
- Many new best practices involving new data don't work very well—or at all—without business metadata, including self-service, data prep, analytics, and data visualization.
- Metadata is the golden thread that stitches together big data architectures and practices, especially in the multiplatform data ecosystems typical of modern data warehousing and customer relationship management (CRM).

NUMBER FOUR

SELECT A PLATFORM OF INTEGRATED DATA MANAGEMENT TOOLS FOR SIMPLIFIED GOVERNANCE VIA MODERN SOLUTION DESIGNS

One of the strongest trends in data management (among both tool providers and tool users) is the movement toward multiple tool types, all integrated into a single platform from a single provider. An integrated platform includes tools for the leading disciplines of data management, such as data integration, data quality, data profiling, metadata management, master data management, event processing, and federation. The integrated toolset should also include recent additions (such as self-service, data prep, and ad hoc exploration) as well as support new big data platforms, formats, and sources.

Note that the collected tools and functions are not a mere suite; instead, they are integrated at appropriate points. For example, in the development environment they share metadata, profiles, business rules, and some development artifacts. In the deployment environment, all functions of all tools can be called from a single, seamless data management workflow.

An integrated tool platform offers advantages for data governance:

Consistent data standards. Creating and adhering to data governance polices for the standards of data (both new and old) is more likely to succeed when most data management solutions are created atop an integrated tool platform. This is simply because there's one development and deployment environment to govern (or fewer of them) instead of a plague of ancillary tools or so-called best-of-breed tools.

Quality metadata, shared broadly for complete views of data. The feature-rich central metadata repository typical of integrated platforms helps to automate some DG tasks, such as inventorying data to be governed and collaborating over data to develop DG policies (especially data quality metrics). Shared, centralized metadata is also key to reconciling and mapping metadata from disparate sources and destinations, both old and new, whether the mappings occur a priori (as in older design paradigms), ad hoc (at runtime, as in new big data practices), or a posteriori (as future needs arise).

Modern solution designs. The reason some tools from vendors and from open source have evolved toward the integrated platform is because it's what savvy users want. For example, for years most data management solutions designed by data management professionals consisted of dozens of individual jobs and routines, sometimes from multiple diverse tools. These were haphazardly stitched together into a solution via scheduling tools. Nowadays, sophisticated data management professionals want to design and architect fewer but more complex data management solutions. In such a modern design, a workflow or data flow at the heart of a single integrated solution in deployment calls multiple tool functions at runtime in a controlled sequence.

Users adopt this approach to gain an optimizable design that has reliable interoperability in production; plus, it allows them to freely mix diverse functions (integration, quality, federation, messaging) to a degree that's unprecedented. Yet they also find that the "big picture" of data and its management seen through a modern design is conducive to the consistent data standards and compliant use of data that's required of data governance.

NUMBER FIVE

CONSIDER HADOOP FOR PERSISTING AND PROCESSING NEW BIG DATA SOURCES, BUT BEWARE OF ITS GOVERNANCE CHALLENGES

Although Hadoop recently had its 10th birthday—and it improves almost daily—it's still a relatively new data platform that has yet to fully embrace data management functionality. Even so, a recent TDWI survey indicates that the number of deployed Hadoop clusters is up 60 percent over two years.[2] Adoption is accelerating because Hadoop is uniquely suited to managing a wide range of new data, and it provides linear scalability for the processing of this data for analytics and operational purposes.

[2] See the discussion around Figure 4 in *TDWI Best Practices Report: Hadoop for the Enterprise* (2015), online at www.tdwi.org/bpreports.

For example, TDWI has found data warehouse professionals enabling multiple use cases on a single Hadoop cluster. To start, Hadoop becomes the new data landing and staging area. Next, they move large collections of extracted source data to Hadoop and execute algorithmic analytics against it. Other pieces of the data warehouse architecture are likewise migrated to Hadoop, such as data marts, operational data stores, and complete customer views. Along the way, the data warehouse team starts integrating big data from new sources into Hadoop and onward into a DW.

Despite its usefulness, the current state of Hadoop presents a few challenges to governing new big data (or any data):

Data replication. The Hadoop Distributed File System (HDFS) automatically (and aggressively) replicates data across its many nodes as part of its high availability strategy. This exacerbates tasks related to DG, such as tracking data lineage and identifying data redundancy. There's a need for smart tools that can understand the semantics of data and (as automatically as possible) catalog Hadoop data, recognize sensitive data managed on Hadoop, and record data's lineage prior to and within Hadoop.

Metadata and its management. Although critical to success, these are still weak on Hadoop. This leaves metadata-driven tools and best practices hamstrung, ranging from queries to data integration to correlating Hadoop data with enterprise data. As mentioned, metadata enables DG tasks such as data inventories, usage audits, and collaboration via data. Luckily, a number of tools from the vendor and open source communities can fill the void.

Nevertheless, keep in mind that metadata is not a mandate on Hadoop. Recall that Hadoop was designed for server logs and other Web data, then processed algorithmically with little or no need for metadata management. This differs sharply from the metadata-driven, set-based operations (often enabled by SQL and/or OLAP) regularly executed on relational systems. As more people (particularly warehousing and analytics specialists) want to execute relational operations on Hadoop (especially for exploration and data prep), better support for metadata and light relational functions on Hadoop become more important to that new audience.

Security. Open source Apache Hadoop supports Kerberos well. However, mature IT organizations want integration with their directory-based systems, and some also want data protection (encryption, masking, tokenization). Again, several vendors are developing tools to address this demand, as is the open source community.

Data store organization. Defining data volumes and other standard ways of organizing data collections are weak on Hadoop today. This inhibits basic database design best practices that isolate data for access privileges (especially DG tasks such as security or privacy) or for performance optimization.

NUMBER SIX

ENSURE THAT NEW BIG DATA HAS THE INFRASTRUCTURE AND GOVERNANCE IT NEEDS TO SUCCESSFULLY MIGRATE ACROSS MULTIPLATFORM DATA ECOSYSTEMS

Back-end data systems are trending toward multiplatform ecosystems consisting of multiple instances of relational DBMSs, including both older brands and new brands based on columns, appliances, and clouds. Hadoop is also emerging as a valuable data platform in these diverse, hybrid ecosystems. Examples of such ecosystems in enterprises today include the modern data warehouse environment and the multichannel customer relationship environment.[3]

Why are so many types of data platforms needed? One reason is that new big data is highly diverse in terms of its structure and how it should be processed. Thus each form of new data can require one data platform that is optimal for its storage and management, whereas that same data may move to another platform for initial processing (e.g., SQL-based exploration) and still another for further processing (e.g., data mining, text mining, graph).

Furthermore, when technical users introduce a new data platform, they usually migrate data to "load balance" storage and processing. In that context, adding Hadoop is disruptive to a data ecosystem's overall architecture and cross-platform data flow. Hadoop is well worth the disruption because of the new data types it can manage and the new analytics and operational processes it can support. Finally, a number of old and new best practices inherently move data around—namely, data exploration, data prep, federated queries, data warehousing, analytics, and agile development methods.

As you can see, new big data tends to migrate a lot in a modern multiplatform ecosystem. On the one hand, the complex architecture and numerous moving parts of these ecosystems can be daunting and somewhat risky for some users. On the other hand, technical users are coping with the complexity just fine, and business people are happy with the new insights they receive from exploring and analyzing new big data.

Given that complex multiplatform data environments are the new norm, users need to build substantial data integration and data quality infrastructure for each ecosystem. If the infrastructure focuses on metadata-driven technical processes, the following benefits will result:

Broad connectivity and access support data integration and quality across many platforms of diverse types and vintages, thereby enabling the continuous data migration and adaptation to the changes required of modern data ecosystems.

[3] For more details about multiplatform data environments, see *TDWI Best Practices Report: Evolving Data Warehouse Architectures* (2014), online at www.tdwi.org/bpreports.

Data lineage, auditing, cataloging, and categorization are enabled by a shared central metadata repository to identify sensitive data and record its movement, thereby enabling data governance that is automated and credible with traditional data, new big data, and Hadoop.

Guided migration is supported by multiple metadata bridges and interfaces, such that metadata developed in one environment (e.g., data integration tools or reporting-tool semantic layers) can be shared with another environment (e.g., Hadoop or discovery tools). This provides useful automation, productivity, governance, and technical standards for the data migrations that are common in today's complex, multiplatform data ecosystems.

ABOUT OUR SPONSOR

Talend.com

Talend is a next-generation leader in cloud and big data integration solutions that helps companies become data driven by making data more accessible, improving its quality, and quickly moving enterprise information where it's needed for real-time decision making. By simplifying big data through these steps, Talend enables companies to act with insight based on accurate, real-time information about their business, customers, and industry. Talend's innovative open-source solutions quickly and efficiently collect, prepare, and combine data from a wide variety of sources, allowing companies to optimize it for virtually any aspect of their business. Talend (NASDAQ: TLND) is headquartered in Redwood City, California. For more information, please visit www.talend.com or follow the company on Twitter: @Talend.

ABOUT THE AUTHOR

Philip Russom is TDWI's senior research director 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 600 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 BI and DW 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.