TDWI RESEARCH

2015–2016

TDWI BENCHMARK GUIDE

TDWI Hadoop Readiness Guide

Interpreting Your Assessment Score

By Philip Russom and Fern Halper

tdwi.org

tdwi
Advancing all things data.

Research Sponsors

Cloudera, Inc.

IBM

MapR Technologies

MarkLogic

Teradata

TDWI BENCHMARK GUIDE 2015–2016

TDWI Hadoop Readiness Guide

Interpreting Your Assessment Score

By Philip Russom and Fern Halper

Table of Contents

About the Authors

PHILIP RUSSOM is a well-known figure in data warehousing and business intelligence (BI), having published over 500 research reports, magazine articles, opinion columns, speeches, Webinars, and more. Today, as the director of TDWI Research for data management, he oversees many of the company's research-oriented publications, services, and events. 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 contributing editor with leading IT magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him by e-mail (prussom@tdwi.org), on Twitter (twitter.com/prussom), and on LinkedIn (linkedin.com/in/philiprussom).

FERN HALPER is director of TDWI Research for advanced analytics, focusing on predictive analytics, social media analysis, text analytics, cloud computing, and other "big data" analytics approaches. She has more than 20 years of experience in data and business analysis, and she has published numerous articles on data mining and information technology. Halper is co-author of "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, service management, and big data. She has been a partner at industry analyst firm Hurwitz & Associates and a lead analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her by e-mail (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/pub/fern-halper/2/491/63).

About TDWI Research

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

Sponsors

Cloudera, IBM, MapR Technologies, MarkLogic, and Teradata sponsored the research for this *TDWI Hadoop Readiness Guide* and its accompanying interactive assessment tool.

Foreword from the Authors

We intended this *Readiness Guide* to help you, the reader, understand Hadoop and the many critical success factors that affect the success of its implementation and use. Furthermore, this guide accompanies TDWI's Hadoop Readiness Assessment Tool, which is a browser-based questionnaire (resembling a survey) that asks how prepared you and your organization are to get full value from Hadoop. When you complete the online questionnaire, the assessment tool immediately shows you scores that quantify your readiness for Hadoop.

This guide provides a primer on Hadoop, an explanation of the Readiness Model that your assessment scores are based on, and tips for interpreting your assessment scores. Thus, it's best that you read this guide before taking the assessment so you are prepared to interpret the scores displayed at the end of the questionnaire. Even so, the guide and the tool both make sense standalone so you can consume them in either order.

The Reason for a Hadoop Readiness Assessment

We are operating under the assumption that people taking the online assessment are contemplating an implementation of Hadoop. Some have zero experience with Hadoop while others have completed a proof-of-concept project with Hadoop but need additional guidance before proceeding with a production implementation. Indeed, at TDWI we regularly communicate with a wide range of people and work with organizations that are considering Hadoop as part of their solutions for big data, advanced analytics, data warehouse modernization, data exploration, and so on. They all have questions about where to focus their best efforts with Hadoop, and the guide and tool for Hadoop Readiness Assessment provide answers for many of these questions.

The Value of a Hadoop Readiness Assessment

The success factors built into the Readiness Model can teach you the most common and most fundamental areas where you need to prepare. Most preparations should be completed before implementing Hadoop, though some can be executed concurrently with an implementation, as explained later. Note that the online assessment tool—when it displays your scores—will also display an average of all other people who took the assessment. That way you can look at the state of your readiness in isolation, or you can add a dimension to that knowledge by comparing your strengths and weaknesses to those of other organizations.

Thank you for reading this *Readiness Guide* and using the Hadoop Readiness Assessment Tool. We sincerely hope you will find both useful.

Philip Russom, Director for Data Management, TDWI Research

Fern Halper, Director for Advanced Analytics, TDWI Research

Hadoop Primer

Before discussing users' readiness for Hadoop, let's step back and define Hadoop for readers who are new to it, as well as define terms so we're communicating accurately.

Defining Hadoop and the Hadoop Ecosystem

Hadoop is a broad ecosystem. The Hadoop ecosystem includes a family of open source software (OSS) Hadoop platforms and tools under the Hadoop brand, distributions of the Hadoop family provided by software vendors, and a growing list of vendor products that interoperate with Hadoop and the data managed by it.

Today, when people say "Hadoop" they usually mean the family of open source products, but at other times they mean the entire, extended ecosystem. This makes sense because users typically deploy many Hadoop products in an integrated fashion, along with vendor products from outside open source.

Hadoop is a family of multiple products and technologies. *Hadoop* is the brand name that the Apache Software Foundation (ASF) and its open source community have given to a family of related open source products and technologies. The Hadoop family includes several products, such as the Hadoop Distributed File System (HDFS), MapReduce, Spark, Hive, HBase, Pig, Mahout, YARN, Avro, Chukwa, and ZooKeeper.[1]

Hadoop's roots are in open source, but it's available from vendors, too. Apache Hadoop is an OSS project administered by the ASF. The Hadoop family is available as open source from ASF as well as from several software vendors. Hadoop and other OSS, for example, can be downloaded freely from the ASF at www.apache.org. Furthermore, a few software vendors offer Hadoop distributions (or "distros") that package the Hadoop family, sometimes with additional tools and features developed by the vendor, typically for administration, security, and other value-adding features. Some packages are Hadoop appliances that deploy quickly and integrate easily with pre-existing relational platforms. These same vendors also offer additional paid support, maintenance, professional services, and additional tools.

Many vendor products now integrate with Hadoop. For example, most mature tools for reporting, analytics, and data integration now have interfaces to various Hadoop tools and platform layers. Furthermore, a few vendors sell new tools that were built specifically for the Hadoop environment. When we pull together Apache Hadoop, vendor distros, mature tools that now interface with Hadoop, and tools purpose-built for Hadoop, we then see a broad Hadoop ecosystem that grows, improves, and offers additional functionality almost daily.

Hadoop enables computational analytics with massive, diverse data sets. The Apache Hadoop software library is a framework that enables the distributed processing of large data sets across clusters of computers, each supporting local computation and storage. Hadoop has a strong reputation for cost-effective, linear scalability with the storage of multi-petabyte data sets—but it does so economically on commodity-priced hardware and low-cost OSS. Big data aside, Hadoop is also a powerful computational platform supporting a wide range of advanced analytic techniques that operate on massive data sets. It is Hadoop's combined support for both diverse big data storage and advanced analytic processing that makes it so compelling to a wide range of organizations.

[1] Readers may wish to read other TDWI reports about Hadoop, especially the TDWI Best Practices Reports *Integrating Hadoop into Business Intelligence and Data Warehousing* and *Managing Big Data*, available for free download at www.tdwi.org/bpreports.

Trends in Evolving Data Management Environments

A number of trends are wending their way through the worlds of IT and data management, trends that point to Hadoop as a solution for problems and an enabler of opportunities.

Organizations need scalability for all data. Many organizations are under growing pressure to scale all enterprise data, whether it is new big data or traditional enterprise data. Though still new, Hadoop has proved its ability to scale linearly with diverse data in petabyte-scale volumes.

Big data is forcing enterprises to rethink the economics of data management. Hadoop isn't free, although uninitiated people sometimes think it is. TDWI, for example, regularly finds users who've paid for a vendor distro to get its valuable support, maintenance, and additional tools. As Hadoop clusters grow, hardware costs and payroll for administration and development do, too. Even so, users tell TDWI that Hadoop is relatively affordable compared to other enterprise platforms for big data management and analytics.

Existing data platforms need greater capacity and life span. A common use case is to extend a data warehouse (DW) by integrating Hadoop into the DW environment. Likewise, Hadoop can extend systems for content management and data archives. In these configurations, Hadoop doesn't replace existing systems. Instead, it offloads some data sets and workloads, expands capacity, and injects new functionality. In turn, this gives existing systems a longer life span, usually at a low cost, which means more value for the enterprise over the long haul.

Organizations want to leverage big data for business value. Never be content to merely capture big data and manage it as a cost center. Instead, organizations should actively explore and process big data to get maximum organizational advantage, often to achieve better customer intelligence, process improvements, fraud detection, and so on. Thus, Hadoop is becoming a preferred platform for collecting new and big data prior to leveraging it.

Firms are eager to compete on analytics. Hadoop isn't just a data platform for managing large data sets. It's also a computational platform that enables advanced forms of analytics, such as those based on data mining, statistical analysis, text analytics, graph, machine learning, and ad hoc algorithmic approaches. This is why Hadoop appeals to for-profit corporations that now seek to compete on analytics. Other organizations apply Hadoop's analytic power to determining the root cause of customer churn, modernizing actuarial calculations, improving patient outcomes in healthcare, and so on.

Organizations need to finally get value from multi-structured data. Early adopters have shown that Hadoop excels at storing, managing, and processing unstructured data (e.g., human language text), semi-structured data (XML and JSON documents), and data with evolving schema (some sensor, log, and social data). Hadoop can make the management and analysis of this diverse data more affordable, scalable, and actionable.

Hadoop's accelerating adoption indicates its value. A recent TDWI survey shows that the number of Hadoop clusters in production is up 60% in two years.[2] Almost half of respondents have new Hadoop clusters in development that will come online within 12 months. At this rate, 60% of users surveyed will have Hadoop in production by 2016, which is a giant step forward.

[2] See Figure 4 in the TDWI Best Practices Report *Hadoop for the Enterprise*, available at tdwi.org/bpreports.

How Hadoop Is Being Used by Organizations Today

A recent TDWI survey asked: "In your perception, what would be the most useful applications of Hadoop if your organization were to implement it?"[3] Data warehousing (DW) and business intelligence (BI) use cases were by far the most common responses to the question. This is no surprise because DW/BI uses cases for Hadoop are well established. However, the prominence of non-DW/BI applications in the survey (e.g., archiving, content management, and operational applications) shows that these are emerging and will become more common. TDWI believes this is a sign that Hadoop usage is diversifying broadly across and within mainstream enterprises.

Data warehouse extensions. Among TDWI members, Hadoop regularly appears as a complementary extension of a data warehouse when warehouse data that doesn't necessarily require the warehouse is migrated to Hadoop. A similar extension is where data staging and data landing functions are migrated to Hadoop. "Fork-lifting" operational data stores to Hadoop is a trend that TDWI has just started seeing.[4]

Analytics and BI. Some of the hottest BI user practices of recent years involve data exploration and discovery, which are critical to learning new facts about a business, as well as getting to know new big data and its potential business value. To enable the broadest possible exploration, some users are collocating numerous large data sets on Hadoop. Data exploration is usually the first step in an analytic project, so it's a fortuitous coincidence that Hadoop is also a capable computational platform and sandbox for advanced analytics. The trend with analytics on Hadoop is toward advanced forms of analytics, such as those based on machine learning, text analytics, graph, statistical analyses, and real-time analytics or event processing.

Lakes and hubs. Data lakes and enterprise data hubs are two of the fastest-growing practices on Hadoop today. Both involve loading multiple massive data sets into Hadoop (easily reaching petabyte scale) with little or no preparation of the data. That way data ingestion is fast, simple, and cheap. To make up for minimal *a priori* data preparation, both lakes and hubs usually rely on post-storage data prep and data federation or virtualization techniques to model and transform data on the fly, on an as-needed basis. This gives analytics the agile ability to repurpose data (at analysis runtime) for open-ended exploration, discovery, analysis, and visualization.

Data archiving. For legal, audit, and compliance reasons, many corporations and other organizations are modernizing their enterprise data archiving facilities. Users are finding that Hadoop has favorable economics and scalability for modern active archives, whether involving non-traditional data (Web, machine, sensor, social) or traditional enterprise data.[5]

Content management. Use cases for Hadoop with content, document, and records management (plus similar practices, such as e-mail archiving) are just now emerging.

Despite the established use cases just described, the current state of Hadoop has weaknesses or omissions that make its use challenging. For example, Hadoop is not a database management system (DBMS), so it lacks DBMS functions for schema and metadata management, indexing, transaction processing, ANSI SQL, granular security, and so on. Luckily, Hadoop gets better almost daily, and vendor products can compensate for many of these challenges.

[3] See Figure 1 in the TDWI Best Practices Report *Hadoop for the Enterprise*, available at tdwi.org/bpreports.

[4] Hadoop's many roles in DW and BI environments are described in the TDWI Best Practices Report *Integrating Hadoop into Business Intelligence and Data Warehousing*, available at tdwi.org/bpreports.

[5] For a description of modern data archive techniques, see *TDWI Checklist Report: Active Data Archiving for Big Data, Compliance, and Analytics*, available at tdwi.org/checklists.

The TDWI Readiness Model for Hadoop

An organization's readiness for Hadoop is not a single state held by a single entity. Corporations, government agencies, educational institutions, healthcare providers, and other types of organizations are complex in that they have multiple departments, lines of business, and teams for various business and technology functions. Each function can be at a different state of readiness for Hadoop, and each function can affect the success or failure of Hadoop programs.

Because of the slight complexity of the average organization, TDWI's Readiness Model for Hadoop assesses Hadoop readiness across five dimensions, most of which map to specific business or technical functions. (See the five dimensions across the top of the Hadoop Readiness Model illustrated in Figure 1.) Of course, within each function there are multiple processes, team structures, levels of experience, and so on that can affect Hadoop's success; these are represented in the model as metrics. (See the metrics listed below each dimension in Figure 1.)

Essentially, the readiness model described here is a compact catalog of success factors for completing an effective implementation of Hadoop. Note that the catalog is compact for a reason. Based on conversations with many users, consultants, and vendors, we have identified the most fundamental success factors relative to readiness and related issues like urgency. One goal is to shorten the assessment tool's questionnaire (which is based on the Readiness Model), so as many users as possible can complete the questionnaire. However, the primary goal is to focus users and their organization on the highest priorities, namely the most common and the most fundamental critical success factors.

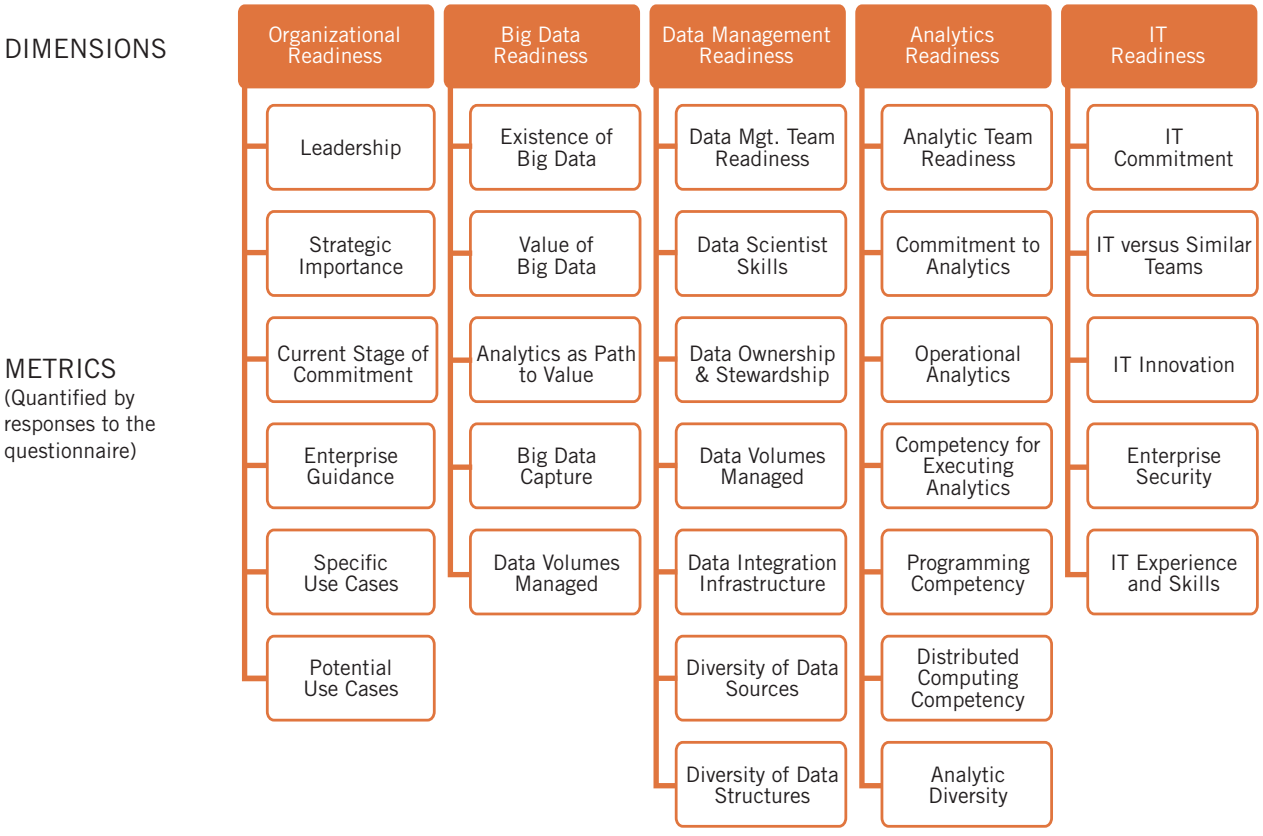

Figure 1. *The Hadoop Readiness Model, consisting of a hierarchy of dimensions and metrics.*

Dimensions and Metrics for Hadoop Readiness

In TDWI's online Hadoop Readiness Assessment Tool, there are one or more questions per metric from the Readiness Model. With multiple metrics per dimension, there are several questions per dimension. In other words, the online tool itself and the scores it displays are based on the hierarchical structure of dimensions and metrics, as defined in the Hadoop Readiness Model found in Figure 1.

Here's a general description of each dimension with a few examples of the questions that the assessment tool asks per dimension and its metrics in order to quantify Hadoop readiness.

Organizational readiness. This dimension concerns three critical success areas: (1) upper management's commitment and vision for Hadoop; (2) the presence of data governance programs and other enterprise programs that can guide Hadoop (and other data-driven projects, too); and (3) whether business-driven use cases for Hadoop have been identified and agreed upon.

Big data readiness. Questions in the online assessment tool test the existence of big data and whether the organization has determined how to get business value from Hadoop and the big data it manages. Questions also ask about a few details of managing big data on Hadoop.

Data management readiness. The assessment tool tests whether a mature data management team and infrastructure already exist and can be assigned to Hadoop, big data, and analytics. Most of the questions of this dimension collect information about the skills and experience of the data management team, because experience with large volumes, exotic data structures, and diverse sources all indicate good readiness for the data types Hadoop manages.

Analytics readiness. Hadoop's ability to scale to petabyte-size big data is legendary, but most users' goal is to build a series of advanced analytic applications atop Hadoop. Accordingly, an organization's commitment to analytics (by business and technology teams) and its prior experience with analytics are important metrics for getting full value from Hadoop.

IT readiness. The assessment tool tests whether an IT team (either centralized or departmental) or some other technical team (e.g., for data warehousing, BI, analytics) has committed to owning, deploying, and maintaining Hadoop. To further quantify readiness, tool questions collect information about IT's inclination toward innovation, general experience, and skills with areas where Hadoop is still maturing, such as security.

How the Readiness Assessment Tool Quantifies Metrics and Dimensions

When a user selects an answer to a question in the Readiness Assessment Tool, the score for that dimension is incremented. For most questions, the multiple-choice answer that a respondent selects determines the score for that question.

When you take the online assessment, resist the urge to inflate your score by answering the questions a certain way. For your assessment to be accurate and useful for your Hadoop planning, you should answer all questions as accurately and honestly as you can.

When a user completes all the questions of a dimension and leaves that dimension's page, a score for that dimension is summed from points earned per question. The greatest score for each single dimension is 20. Multiplying 20 by the 5 dimensions yields 100 as the greatest possible score overall.

At the end of a respondent's session, the Readiness Assessment Tool displays that respondent's scores per dimension and overall score, plus the average dimension and overall scores of all respondents.

That way you have a context for determining whether your organization is ahead of or behind the curve in the aggregate.

If a respondent is today completely prepared for Hadoop, his or her score might tally to 100, but that's a rare person, and most overall assessment scores will fall between 60 and 80. An overall score of 50 is a reliable watershed benchmark; a score at or above 50 indicates that the organization meets the basic readiness requirements for a successful Hadoop implementation.

Further preparation can be successfully executed concurrently during the implementation. Below that, there are most likely improvements that should be made to use-case commitments, goals for business value, technical skills, and technical infrastructure before a Hadoop implementation commences. Even so, there are many ways to interpret assessment scores, as explained in the next section of this guide.

Average Readiness Scores across All Respondents

Note that at the time this guide was written and published, the online Hadoop Readiness Assessment Tool had not yet gone live and collected response data from people taking the assessment. Consequently, the following examples of scores are based on what we at TDWI think will happen based on prior experience with other assessment tools.

As mentioned earlier, the online tool will show your scores after you complete your session, as well as an average of scores across all assessment respondents. That way you can benchmark your scores against those of your distant peers at organizations that are also contemplating an investment in Hadoop.

The average scores could go in a variety of directions and will evolve over time as more people complete the Hadoop Readiness Assessment. Note that all data sets have one or more biases based on the diversity of the data's sources and other factors. Survey results, for example, are often biased by the population of people who take the survey. With the Hadoop Readiness Assessment questionnaire, TDWI thinks there are certain profiles of organizations that may dominate the respondent population and therefore influence the average scores:

People from mature and sophisticated organizations will probably take the assessment. If these people dominate, then the average scores could look like those illustrated in the "radar" chart of Figure 2. This reading reveals a high competency across all five dimensions of the Readiness Model, as seen in the fairly symmetrical pentagonal shape inside the chart. Because these organizations are equally good in all dimensions, the pentagon's edges are near the outer edge of the radar chart (a position that denotes strength, as opposed to weakness nearer the center). These organizations can handle anything and should proceed with Hadoop, assuming communication across the dimensions has resulted in the identification of early-phase, business-driven use cases.

The respondent population might be dominated by data management professionals (or specifically BI, DW, and analytics professionals). If that happens, their organizations will score very high in data management readiness and score fairly well in related dimensions, such as big data and analytics. (See Figure 3.) Organizations of this profile are well positioned for analytic success with Hadoop, although they may need to secure deeper commitments from IT and business management to ensure proper support, funding, and sponsorship.

Respondents who work in the same industry tend to have similar readiness profiles. For example, TDWI's experience with manufacturers suggests that most organizations in that industry are capable of implementing Hadoop. This is due to deep data management and IT experience—bolstered by solid organizational support and guidance—but despite inexperience with big data and analytics, as illustrated in Figure 4. When you complete the online questionnaire, the Hadoop Readiness Assessment Tool will show you the average for your industry, so you have an added dimension of context.

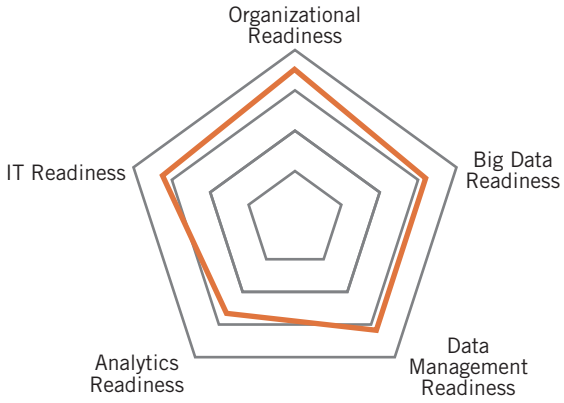

Figure 2. *Possible scores, averaged across all assessment tool respondents, when competency is high across all dimensions for a large percentage of respondents.*

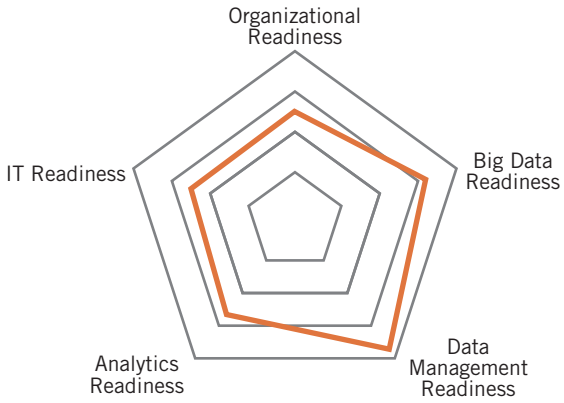

Figure 3. *Possible scores, averaged across all assessment tool respondents, when data management professionals dominate the respondent population.*

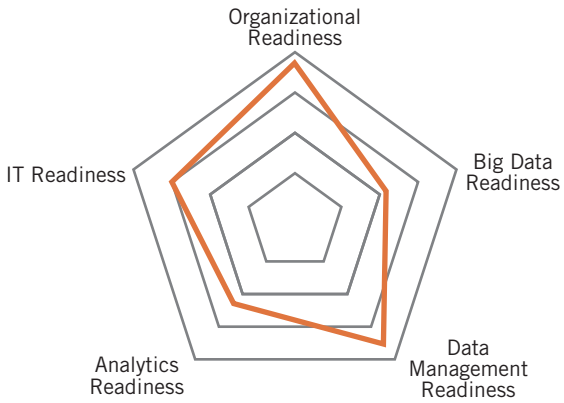

Figure 4. *Possible scores averaged across assessment tool respondents who work in the manufacturing industry.*

Interpreting the States of Hadoop Readiness

To help you interpret your assessment results (as returned by the TDWI Hadoop Readiness Assessment Tool), here are five common scenarios for Hadoop readiness along with background, interpretations, and recommendations for each.

Readiness Scenario No. 1: Minimal Organizational Support

Everyone pays lip service to the fact that all IT programs (especially new ones) have a better chance of meeting business goals when the program has solid executive support, ownership, and funding. In a similar issue, TDWI survey results and users speaking at TDWI events all agree that having a data governance program in place greatly increases the likelihood of success for data-driven business and technology programs. Likewise, the business payback of an IT platform is less risky when there are specific applications and use cases defined up front by managers from both technology and business. Hence, the absence or weakness of any of these definitions of organizational readiness can raise the risks associated with Hadoop and other programs, as seen in the scores of a hypothetical organization charted in Figure 5.

Strengths: The organization represented in Figure 5 is lucky to have strong technology dimensions, namely those for big data, data management, and IT. Given these strengths, the slight weakness in analytics is probably due to inexperience or management negligence, both of which are easy to cure.

Weaknesses: Although big data has arrived in this organization, business management isn't yet convinced that it has value, or they haven't studied big data as an enterprise asset to learn where the leverage points are. The weakness in analytics is also most likely due to management being behind in trends for modern data-driven business practices.

Recommendations: Someone needs to sell C-level executives on the value of analytics and big data so they'll bestow ample resources on Hadoop. We talk of "modernizing" various types of IT systems, but many organizations also need to modernize business management so it can adopt new operational and competitive business models based on broad insights from big data and advanced analytics.

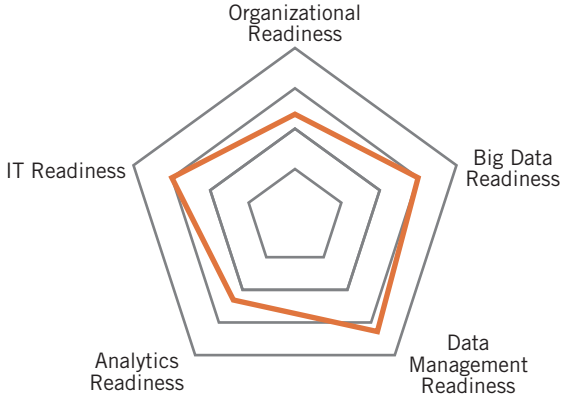

Figure 5. *Possible individual scores when organizational management isn't analytic enough.*

Readiness Scenario No. 2: Big Data Has Not Yet Arrived or Accumulated

It's natural that organizations don't feel the urgency for leveraging big data until it starts arriving and accumulating in appreciable quantities. It takes a certain critical mass of big data volume before exploring and profiling it can yield accurate appraisals of big data's unique value to a specific organization. Therefore, even when users see that big data is coming, their sense of urgency and their plan for leveraging it can be limited, as seen in the possible scores charted in Figure 6.

Strengths: On the upside, the organization represented in Figure 6 is already committed to capturing, governing, and analyzing big data, although it's just arriving. In addition, it already has solid competencies in data management and analytics. Given the organizational will and the data infrastructure in place, this organization should proceed with Hadoop.

Weaknesses: On the downside, the slight weakness in IT readiness is probably due to a lack of experience with distributed server clusters, open source software, and programming or scripting languages required by Hadoop (R, Java, Python, etc.). Training is available from a number of sources and many IT teams have a record of learning these quickly, so this weakness is easily remedied.

Recommendations: Priority should go to beefing up IT and similar teams to learn the skills, acquire the tools, and deploy the server infrastructure needed for Hadoop. Given that big data is just arriving, there is ample time (and organizational support) to improve IT, start a Hadoop program, and thus be positioned to get full value from big data as soon as it accumulates.

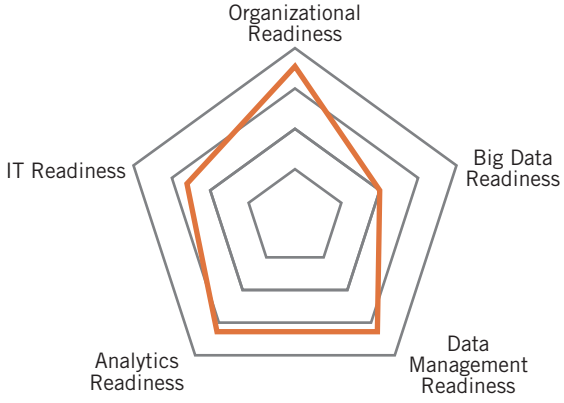

Figure 6. *Possible individual scores when big data is lacking size, diversity, or value.*

Readiness Scenario No. 3: Nascent Data Management Maturity

New or young organizations (whether simply departments or full enterprises) are often so small and simple that minimal IT and data management suffice. However, as the organization grows and emerges from its start-up phases, business managers are under pressure to run the business based on the numbers and make fact-based decisions. That's the point where they suddenly get serious about data management in support of tried-and-true programs for business intelligence (BI), data warehousing (DW), data integration (DI), and analytics. TDWI has seen this maturation pattern many times, and large Internet firms have spoken at TDWI conferences about the moment where data management (both traditional and modern) becomes an imperative. That moment is seen in the scores of Figure 7.

Strengths: In this organization, the IT department is more of a research and development (R&D) team because it is dominated by application developers who program homegrown applications. Thus, IT is very strong and has a mindset that fits Hadoop well.

Weaknesses: This kind of IT team is talented with programming, but its competencies do not extend far into data management. For example, teams of this profile regularly think they are executing analytics when they are just running reports. They think they built a data warehouse when all they built was a simple operational data store.

Recommendations: An IT organization (or R&D team) with this much talent will have no trouble implementing Hadoop. However, getting full analytics value from the big data that Hadoop manages will require them to build a data competency that complements their programming competency. For maximum benefit, the new data management competency center should be staffed broadly with database administrators, BI/DW professionals, data analysts, and data scientists.

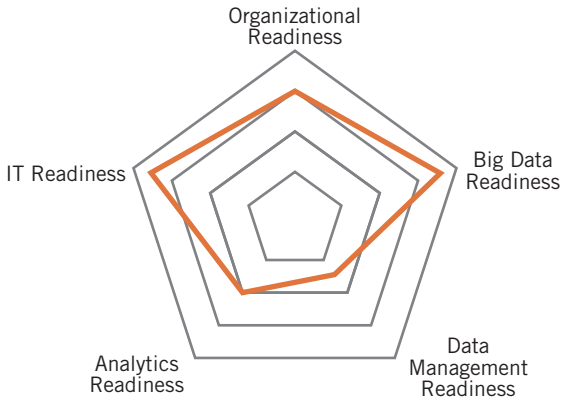

Figure 7. *Possible individual scores when data management is immature.*

Readiness Scenario No. 4: Little or No Advanced Analytics

Online analytic processing (OLAP) continues to be the most common form of analytics in place today, and it won't go away because of its value to users. The catch with OLAP is that scope of discovery enabled by an OLAP tool is inherently limited by fixed data models, whereas advanced analytics usually avoids such limitations by working with raw source data and by building models on the fly during exploration.

With that in mind, today's trend toward analytics can be described as users adopting additional forms of discovery-oriented analytics that complement older OLAP and reporting approaches. In that scenario, many organizations have mature programs for OLAP, reporting, DW, and DI, whereas advanced analytics is still very new or nonexistent, as seen in the scores of Figure 8.

Strengths: This profile of user organization is well positioned to embrace Hadoop due to its deep experience with data management backed up by strong IT and organizational readiness.

Weaknesses: The combinatorial, set-based skills learned from OLAP design do not translate well to the algorithm-driven analytics often performed in Hadoop environments. Even so, experienced BI/DW teams have a track record of success in learning new development skills.

Recommendations: This organization should proceed with Hadoop but simultaneously beef up its competency with advanced forms of analytics. TDWI survey data shows that organizations in this situation prefer to cross-train existing BI/DW staff members in analytics instead of trying to hire from the limited pool of analytics professionals available in today's job market.[6]

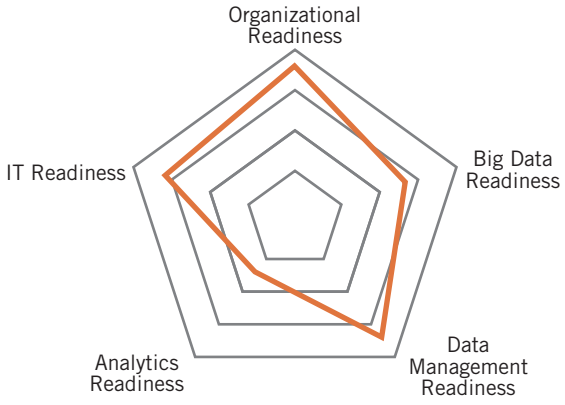

Figure 8. *Possible individual scores when the organization does OLAP but not other analytics.*

[6] See Figure 10 in the TDWI Best Practices Report *Hadoop for the Enterprise*, available at tdwi.org/bpreports.

Readiness Scenario No. 5: Weak IT Ownership or Experience

A technical team needs to commit to owning, deploying, and maintaining Hadoop—but which team? It could be IT, whether IT is a centralized shared service or a pseudo-autonomous departmental team, or the owner could be an applications team or BI/DW team. The trend is toward Hadoop as a shared enterprise asset, provided by central IT, the way that most central IT teams provide networks, storage, and racks of CPUs for multiple applications. TDWI has seen all these owner types succeed with Hadoop. Hence, any owner is far better than no owner, and yet some organizations have trouble establishing ownership, as illustrated in Figure 9.

Strengths: This profile of user organization is clearly data-driven, with good big data, data management, and analytics readiness. Furthermore, business and governance functions are ready to foster and guide Hadoop.

Weaknesses: Alas, IT readiness is low largely due to the lack of an established and committed owner for Hadoop. Other weaknesses may be due to a lack of experience with Hadoop technologies and skills, such as OSS, MPP computing architectures, and programming in Hadoop languages (C, Java, R, Python).

Recommendations: Establish an owner. The technology team that eventually takes ownership of Hadoop may need to improve its skills and infrastructure prior to a Hadoop implementation. Improvements usually include training in Hadoop technologies and acquisitions of hardware appropriate to Hadoop clusters.

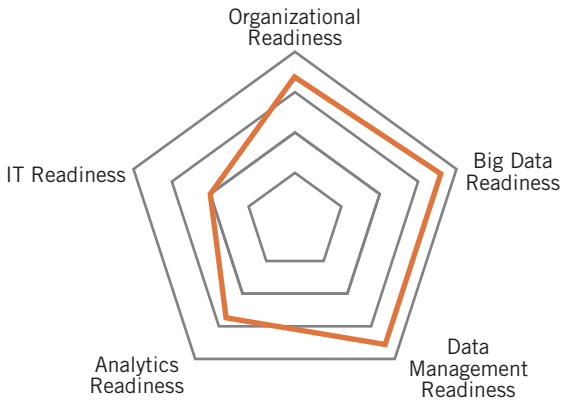

Figure 9. *Possible individual scores when IT isn't ready for Hadoop.*

Summary

The TDWI Hadoop Readiness Assessment provides a quick way for organizations to assess their readiness for Hadoop and compare themselves in an objective way against others with Hadoop initiatives. The assessment is based on the TDWI Hadoop Readiness Framework, which consists of 5 dimensions and 30 questions across these 5 dimensions. Although this assessment serves as a relatively coarse measure of your readiness, we think you will find it useful.

cloudera

Cloudera, Inc.

www.cloudera.com

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, process, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Only Cloudera offers everything needed on a journey to an enterprise data hub, including software for business-critical data challenges such as storage, access, management, analysis, security, and search. As the leading educator of Hadoop professionals, Cloudera has trained over 22,000 individuals worldwide. Over 1,000 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.

IBM business

IBM

www.ibm.com

IBM is the global leader in big data and analytics and offers the most comprehensive solutions for Hadoop and for the Hadoop ecosystem. These solutions include the ODPi-compliant IBM Open Platform (IOP), a free Hadoop distribution, in-Hadoop analytic technologies to accelerate the conversion of data into valuable insight for businesses, and leading-edge Spark-enabled solutions geared at the growing profession of data science and data engineering. IBM's offerings are available in the cloud, enabling customers to quickly scale their capabilities to meet their growing workloads.

MAPR

MapR Technologies

www.mapr.com

MapR delivers on the promise of Hadoop with a proven, enterprise-grade platform that supports a broad set of mission-critical and real-time production uses. MapR brings unprecedented dependability, ease of use, and world-record speed to Hadoop, NoSQL, database, and streaming applications in one unified big data platform. MapR is used by more than 700 customers across financial services, retail, media, healthcare, manufacturing, telecommunications, and government organizations as well as by leading *Fortune* 100 and Web 2.0 companies. Amazon, Cisco, Google, and HP are part of the broad MapR partner ecosystem. Investors include Google Capital, Lightspeed Venture Partners, Mayfield Fund, NEA, Qualcomm Ventures, and Redpoint Ventures.

MarkLogic

MarkLogic

www.marklogic.com

For more than a decade, MarkLogic has delivered a powerful, agile, and trusted enterprise NoSQL database platform that enables organizations to turn all data into valuable and actionable information. Organizations around the world rely on MarkLogic's enterprise-grade technology to power the new generation of information applications. MarkLogic is headquartered in Silicon Valley and has offices throughout the U.S., Europe, Asia, and Australia. For more information, please visit www.marklogic.com.

TERADATA

Teradata

www.teradata.com

Teradata helps companies get more value from data than any other company. Our big data analytic solutions, integrated marketing applications, and team of experts can help your company gain a sustainable competitive advantage with data. Teradata helps organizations leverage all their data so they can know more about their customers and business and do more of what's really important. www.teradata.com

TDWI RESEARCH

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

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