research

BEST PRACTICES REPORT

DECEMBER 2016

Data Science and Big Data

Enterprise Paths to Success

By Fern Halper

tdwi
Transforming Data
With Intelligence™

Research Sponsors

IBM

MapR

OpenText

Snowflake Computing

research

BEST PRACTICES REPORT

DECEMBER 2016

Data Science and Big Data

Enterprise Paths to Success

By Fern Halper

Table of Contents

About the Author

FERN HALPER, Ph.D., is vice president and senior 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 has published numerous articles on data mining and information technology. Halper is coauthor of "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, and service management, as well as *Big Data for Dummies*. 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 at fhalper@tdwi.org, @fhalper on Twitter, and on LinkedIn at linkedin.com/in/fbhalper.

About TDWI

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data management industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence, analytics, and data management solutions. TDWI also fosters the advancement of business intelligence, analytics, and data management research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, six major educational conferences, topical educational seminars, role-based training, onsite and online courses, certification, solution provider partnerships, an awards program for best practices, live webinars, resource-filled publications, an in-depth research program, and a comprehensive website: tdwi.org.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies and is supplemented by surveys of business intelligence professionals. To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving business intelligence problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical BI issues. To suggest a topic that meets these requirements, please contact TDWI senior research directors Fern Halper (fhalper@tdwi.org), Philip Russom (prussom@tdwi.org), and David Stodder (dstodder@tdwi.org).

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who agreed to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: James Powell, Lindsay Stares, James Haley, Peter Considine, and Michael Boyda.

Sponsors

IBM, MapR, OpenText, and Snowflake Computing sponsored the research and writing of this research report.

Research Methodology and Demographics

Report Purpose. This report educates organizations in best practices and options for big data and data science. This includes organizational strategies for deploying data science as well as big data technology options and other considerations. The report also examines how organizations are using big data and analytics and gaining value.

Terminology. *Big data* refers to the capability to manage large volumes of disparate and multistructured data at the right speed and within the right time frame to enable analysis and action. *Data science* includes the tools and techniques to help analyze this data, which often involves modeling and programming.

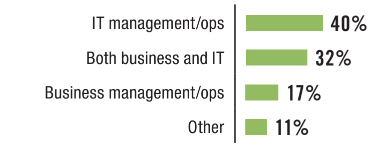
Survey Methodology. In August 2016, TDWI sent an invitation via email to the BI and data professionals in our database, asking them to complete an online survey. The invitation was also posted online and in publications from TDWI and other firms. The survey collected responses from 370 respondents. A total of 296 respondents completed all questions. All responses are valuable and so are included in this report's data sample. This explains why the number of respondents varies per question.

Research Methods. In addition to the survey, TDWI conducted telephone interviews with technical users, business sponsors, and data management experts. TDWI also received briefings from vendors that offer products and services related to big data and analytics.

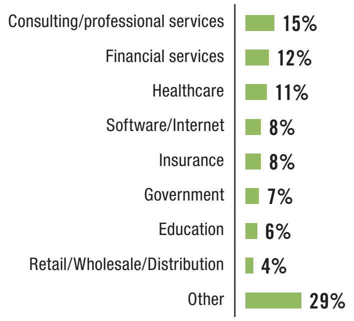
Survey Demographics. The majority of survey respondents are from IT (40%), followed by those who straddle both business and IT (32%) and then other business users (17%). A large percentage identified as data analysts, architects, and consultants. The majority (51%) have 10+ years of experience in BI/analytics and/or data management.

The consulting (15%) and financial services (12%) industries dominate the respondent population, followed by healthcare (11%), software/Internet (8%), insurance (8%), and other industries. Most survey respondents reside in the U.S. (60%) or Europe (15%). Respondents come from enterprises of all sizes.

Position

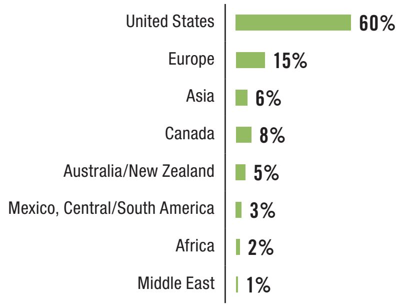


Industry

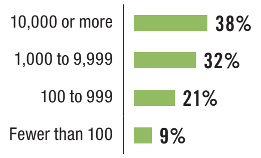


("Other" consists of multiple industries, each represented by less than 4% of respondents.)

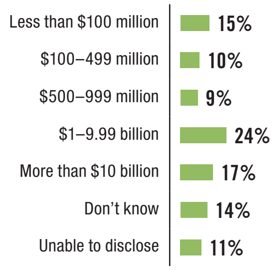
Geography



Number of Employees



Company Size by Revenue



Based on 296 respondents who completed every question in the survey.

Executive Summary

Big data and data science can provide a significant path to value for organizations. These technologies, methodologies, and skills can help organizations gain additional insight about customers and operations; they can help make organizations more efficient, be a new source of revenue, and make organizations more competitive. Although many companies are still analyzing structured data, "newer" data sources such as text data, streaming data, and geospatial data are becoming part of an evolving data landscape. However, businesses often struggle with putting an effective analytics and data science strategy together. Part of the issue is with technology. Additionally, there are organizational challenges that must be addressed, which often include building cultures, hiring the right people, and organizing to execute.

There are many paths to value with big data and data science.

TDWI Research finds that there are many paths to value. On the technology front, our survey respondents are utilizing open source and commercial packages to drive big data value. They are using a mix of on-premises and cloud technologies. They are using data warehouses together with newer technologies such as Hadoop and Spark to manage and process big data. They are deploying appliances, MPP (massively parallel processing) databases, and other solutions to meet their big data management needs. A new ecosystem is evolving to support big data and data science.

On the analytics front, data scientists and others needed to succeed in big data are often hard to find. Survey respondents are using different approaches to build data science skills. They are hiring from the outside as well as trying to grow talent internally. They are often looking to business analysts to become more sophisticated analytically to supplement data science expertise. Some are using a team approach. Many organizations are creating centers of excellence to provide analytics and big data expertise and to disseminate learning. Few are hiring chief analytics officers or chief data scientists. Most look to the VP of analytics or the CIO to help them in their efforts.

This TDWI Best Practices Report examines organizations' experiences with and plans for big data and data science including both technology plans and organizational strategies. It also looks at various big data challenges and how organizations are overcoming them. It examines the importance of new open source models. Finally, it offers recommendations and best practices for successfully implementing big data programs in the organization.

A unique feature of this report is its examination of the characteristics of companies that have actually measured either top-line or bottom-line impact with big data and data science. In other words, it explores how those companies compare against those that haven't measured value.

An Introduction to Big Data and Data Science

Defining Big Data and Data Science

The data and analytics landscape is changing. Although many organizations are still analyzing structured data from their data warehouse, TDWI research indicates organizations have increasing interest in analyzing disparate kinds of data. This includes text data, streaming data, geospatial data, and machine-generated data—to name a few. More organizations are beginning to analyze large volumes of this "new" data. They are modernizing their data infrastructures and platforms to meet this need. They are developing new skills and tooling to support analytics efforts that incorporate this bigger universe of data.

TDWI research indicates increasing interest in analyzing disparate kinds of big data.

The industry around big data and data science is one result of this evolution/revolution. Although the market often uses the terms *big data* and *data science* interchangeably, they are really quite different. Big data refers to the capability to manage large volumes of disparate data at the right speed and within the right time frame to enable analysis and action. Big data is about the three *v*'s—*volume*, *variety*, and *velocity*—and some would add *value*. Organizations are moving toward more hybrid environments to manage this big and multistructured data. This often includes the cloud, Hadoop, and data lakes as well as NoSQL databases and other platforms. Big data analysis frequently requires the use of MPP (massively parallel processing engines), in-memory computing, and other technologies that can handle large quantities of data.

Data science, along with the role of data scientist, in many ways is an outgrowth of the need to analyze big data. Data science is an interdisciplinary field that extracts insights from data. The relatively new name *data scientist* refers to a well-established group of professionals who engage in statistical analysis and exploration: data mining professionals, predictive modeling professionals, etc. Some are part computer scientist, part statistician, part mathematician, and part business analyst. The data scientist (or data science team) also develops new algorithms and applications and analyzes data (often big data) using advanced analytics techniques such as machine learning and natural language processing. For instance, a data scientist might build an application to analyze large amounts of disparate data types to understand fraud or churn. They might build recommendation systems or image recognition systems or systems to optimize operations.

Along with the growth in data science is the rise of *citizen data scientists*. These are the next generation of statistical explorers, sometimes from nontraditional backgrounds, who are variously self-taught, self-starting, self-sufficient, and self-service in orientation. These are often business users or analysts who may not have formal training in statistics or math but perform advanced analytics using some of the easy-to-use advanced analytics software being marketed by analytics vendors.

***Data scientist* is a new name for a well-established group of professionals.**

Another recent trend in big data and data science includes the growing use of open source technologies as part of the big data/data science landscape. These include Hadoop, Spark, and others discussed in this report. Likewise, the cloud is also becoming a big part of the equation as organizations look to scale up their data management and analytics capabilities in the face of big and new kinds of data.

The growth in data and the realization that analytics does provide value is leading organizations down multiple paths in terms of how to execute on their analytics vision. There is not one distinct path to success; rather, organizations are looking to different technologies and business practices to help them get started and grow their efforts.

Data Science Technologies

Although data scientists make use of the existing commercial software and statistical algorithms that are already on the market, when discussing data science, "newer"[1] technologies often come to mind. Some of these much-talked-about technologies include:

Recently, vendors have started to support open source R.

- **NLP (natural language processing).** An important component of data science, NLP involves analyzing, understanding, and generating languages to ultimately enable interfacing with systems using human language rather than computer languages. For text, NLP often uses semantics to parse sentences to understand entities (people, places, things), as well as concepts (words and phrases that indicate a particular idea), themes (groups of co-occurring concepts), or sentiments (positive, negative, neutral). NLP is often part of text mining solutions and is also part of the evolving landscape of cognitive computing and AI (artificial intelligence).
- **Machine learning.** This involves building systems that can learn from data to identify patterns and predict future results with minimal human intervention. The computer learns from examples using either supervised or unsupervised approaches. The algorithms are provided with historical data (multistructured) with known outcomes for training. A test set is then utilized to determine how well the model performs with new data. Machine learning is often used in predictive analytics and utilizes mathematical and computational science approaches as well as statistical approaches.
- **Open source analytics environments such as R.** A language and environment for statistical analysis, R is part of the GNU free software/open source project. It includes data handling and storage facilities, a large set of tools for data analysis (including machine learning and NLP), tools for graphical analysis, and its programming environment. R has been around for several decades (it is similar in many ways to the S language developed at Bell Laboratories in the 1970s) and is widely used as a statistical environment by universities as well as corporations. There is huge community support for R. Recently, vendors have begun to support R, often providing new features and functionality around some of its shortcomings, such as security, along with native connectors and optimizations for their products.
- **Programming languages such as Python and Java.** Python is an interpreted, interactive, object-oriented scripting language now available through the Python Foundation. Like R, it was developed in the 1990s to be an easy-to-read language and has a library for analytics. Developers like its flexibility and its simplicity. Although Hadoop is written in Java, MapReduce applications can be written in Python (and R). Python supports other Hadoop ecosystem projects such as Spark, Storm, Hive, and HBase. Python is often used in developing Web applications. As with R, some vendors are also now providing Python connectors to their products.
- **Operating environments such as Spark.** Spark is an open source big data processing framework that is part of the Apache project. The framework provides processing capabilities for multiple kinds of big data (text, graph, streaming). Spark also offers analytics libraries, including a machine learning library. Theoretically, it is able to process data faster than MapReduce because it processes data in-memory while MapReduce persists back to the disk after a map or reduce action. This is helpful for iterative analysis. Again, some vendors are now providing Spark connectors to their non-Hadoop products.

[1] In this context, "newer" refers to technologies, some decades old, that are receiving renewed interest because of the evolution in data science.

Use Cases and Examples

The use cases for big data are wide and varied. We asked respondents, experts, and vendors how they are using big data and data science. Some of their answers include:

Customer-related analytics in marketing. Marketing is an important area for big data and analytics. As more companies compete through data-driven marketing, sharpening efforts to attract and retain customers requires organizations to handle large volumes of data coming from multiple channels and sources. In previous TDWI research, we've seen that marketing is often one of the first areas to use more advanced analytics. Organizations collect data about customers, including social media data, emails, and clickstream data, that can add insight about the customer journey, determine how to best market and segment customers, and improve their experience. Some advanced analytics use cases that respondents cited in this survey include churn and retention analysis, up-sell, next best offer/action/communication, real-time customer insights, customer sentiment, customer service and satisfaction, customer loyalty, recommendation engines, marketing strategy development, and customer journey analysis.

Marketing is often one of the first areas to make use of more advanced analytics.

Fraud and risk analysis. Many respondents, spanning multiple industries, cite fraud and risk analysis as a use case for big data. For instance, respondents at insurance companies cited the use of big data for risk and fraud detection in claims processing. Insurers are also using machine data from vehicles to determine risk associated with drivers. For example, some might use sensor data to track aggressive driving and hard braking in insured vehicles. In finance, respondents talked about portfolio and investment risk and credit card fraud detection. In healthcare, big data is used to predict the risk of patients developing hospital-acquired infections. In education, it is used to determine the risk of students dropping out of school.

Preventive maintenance in asset management. Many respondents mentioned predictive or preventive maintenance as a use case for big data. Some of these examples make use of sensor data from the Internet of Things (IoT).[2] For instance, a fleet operator might use sensors to collect data from their various trucks. Such data might include the temperature or number of vibrations per second of a particular part or parts. This data can be analyzed to determine what precipitates a part failure or when undue wear and tear is occurring. That information might then be encoded into a set of rules or a model and used to score new data from trucks in order to improve fleet maintenance and operational efficiency.

Patient-related analytics in healthcare and pharmaceuticals. Over the past few years, we've been seeing healthcare as an emerging area for advanced analytics. A number of respondents cited healthcare-related big data use cases, which, in addition to predicting infection risk, include measuring health quality, proactive patient health engagement, population health analysis, trending of health conditions, patient population care pathway optimization, and understanding the patient journey.

Gaming. Gaming analytics is an emerging big data use case. In gaming analytics, big data is analyzed to balance the game as well as to understand player behavior to increase engagement and drive revenue.

[2] For more information on IoT, see *TDWI IoT Readiness Guide* (2016), online at bit.ly/IoTreadiness.

Drivers for Big Data

There are numerous examples of big data and data science and many reasons why the market for big data is growing, but what are the drivers for actual user adoption of big data, big data analytics, and data science? We asked all respondents to score the importance of a number of drivers. On the 5-point scale, 1 was extremely unimportant and 5 was extremely important. Drivers rated as *extremely important* centered on various aspects of the business, such as customer understanding or operational efficiencies (Figure 1).

More accurate business insight is the top driver for big data analytics.

Top drivers focus on better insight. At the top of the list of extremely important drivers were more accurate business insights (57%) and understanding customers (49%). Although more volume of the same data in and of itself doesn't necessarily lead to better insights, more diverse kinds of data and analyses can drive better insight. For instance, social media data can help companies understand customer sentiment, as can text in emails. Claim notes can provide insight into fraud. Telematics data can provide insight into driver habits. Data from beacons can help retailers understand how buyers are moving around a store. Sensor data from railroad cars can provide useful information about wear and tear or possible rail problems. Clearly, respondents are interested in big data's ability to help drive a better understanding of what is happening in their businesses and with customers.

Predicting behavior and business performance are also important. In addition to driving understanding, respondents are also interested in using big data to predict behavior (44%) and to improve business performance and processes (40%). This illustrates that respondents are interested in actually *doing something* with increased insights. Interestingly, although the top driver remains more accurate business insights, drilling into the data by industry indicates that financial services and insurance company respondents put identifying risk and fraud (ranked eighth among all respondents) in the top three extremely important drivers. Healthcare respondents put predicting behavior higher on the list than the overall group of respondents.

Close to a third of those with big data programs cite monetizing analytics as an extremely important driver.

Building apps is low on the list. It is also illuminating to compare the top drivers with those that are lower on the list. For instance, respondents are interested in greater insight and understanding around customers and business processes, but they are less interested in building analytics applications (19% rated extremely important) or monetizing their analytics (22% rated extremely important). This might represent that the market is still relatively new when it comes to big data and analyzing disparate kinds of data. Many respondents are still at the early stages in terms of thinking what they want to do with it. However, early adopters are more likely to be thinking about some way to monetize their analytics. When the data is filtered for just those organizations who already have a big data program underway (about a third of the respondents), a higher percentage of respondents (32%, not shown) cite this as an extremely important driver.

Percentage of respondents ranking the following drivers for big data deployments "extremely important" in their organization or company.

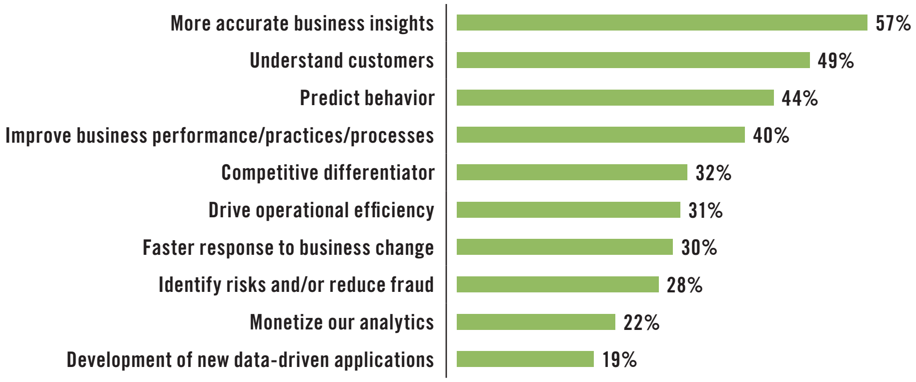


Figure 1. *Based on 352 respondents.*

The State of Big Data and Data Science

Although there has been a lot of market hype and excitement around big data and data science, this does not necessarily mean that it has widely penetrated in most organizations. As previously stated, a little more than a third of our respondents believe they have a big data and analytics program in place now, and another third plan to have one within the year.

We asked respondents where they are now in terms of assembling large volumes of data and using different data types for analytics, as well as where they are with their data management and analytics programs.

Big Data Volumes and Types

Approximately 60% of respondents are managing data in the terabyte range.

Big data includes large volumes of disparate data; in other words, it includes structured data as well as multistructured data such as text data, geospatial data, or streaming sensor data—to name but a few. These "new" kinds of data are driving what big data and data science is all about. We asked respondents about how much data they are collecting and what kinds of data they are using for big data analysis (Figure 2). Even though about a quarter of the respondents are still dealing with gigabytes of data, the majority of respondents appear to be collecting and trying to analyze data in the terabyte range. About 30% are collecting and analyzing data in the 1–10 TB range. An additional 20% are capturing and analyzing data in the 10–100 TB range; and 10% are in the 100–1,000 TB range. A small percentage (7%) are collecting and analyzing petabytes of data. Using data volumes alone would suggest that more than a third of respondents are already dealing with some form of big data, contrary to how respondents self-identified.

How much data does your organization collect and use for analytics?

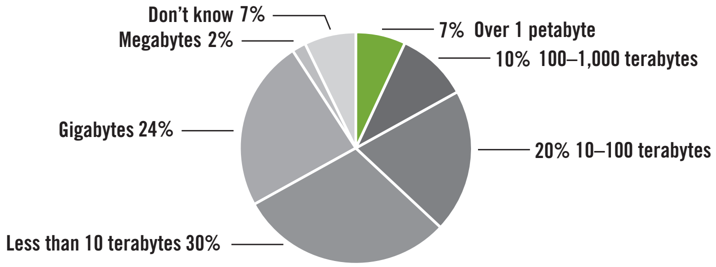


Figure 2. *Based on 370 respondents.*

Respondents have plans for managing new kinds of data.

Much of this data is structured data (Figure 3), although respondents have plans to bring other data types into the fold. We asked respondents what kinds of data they are managing as big data now and what they plan to be managing two years from now. Overall, the largest kind of data is structured data found in databases and data warehouses (82%). This is often traditional data that companies are using for analytics. However, organizations are also managing and analyzing more complex data such as semistructured (45%) and complex hierarchical data (46%), geospatial data (39%), and log data (40%).

Nearly half of organizations with a big data program in place are managing text data now.

Respondents have plans for new kinds of data. That is not to say that respondents are not utilizing other kinds of data. Those who already have a big data program in place are more likely to be managing these disparate kinds of data than those who are planning to have one. For example, 27% of all respondents are managing text data now, but 48% of those who already have a big data program in place are managing that data now. Although 17% of all respondents claim to be managing machine-generated data now, 35% of those that already have a big data program in place are managing machine-generated data. This also reinforces the point that some respondents are already managing large, diverse data sets but don't think of themselves as having a big data program/data science program deployed.

Many respondents plan to bring this newer data into the organization. For instance, 44% of respondents plan to utilize text data, 41% plan to use social media data, and 43% plan to use real-time streaming data in the next two years. Interestingly, 34% of respondents plan to be managing machine-generated data in the next two years. There appears to be well-warranted excitement in the market about these data types. Utilizing disparate data, such as text data or geospatial data, as part of an analysis can improve model accuracy and provide more insight. Streaming data can provide real-time insights. TDWI has seen in past research that those organizations using disparate data types for analytics are more likely to be able to gain measurable value from their analytics efforts. The key is to start using this kind of data.

Which of the following types of data are you managing as big data now? Within two years from now?

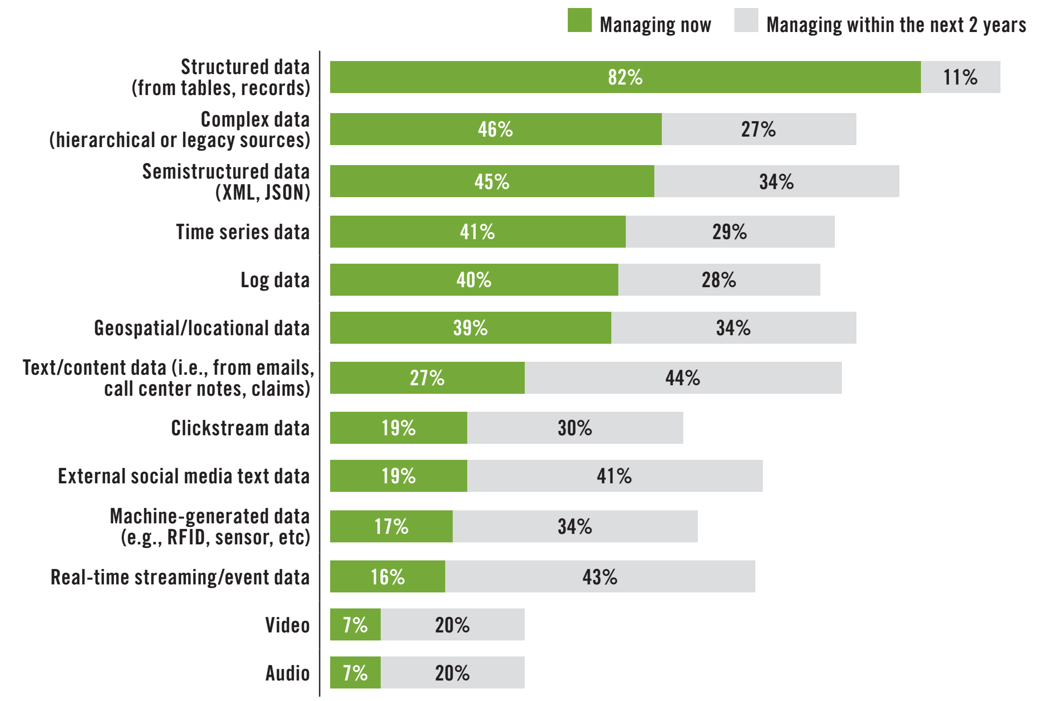


Figure 3. *Ordered by percentage managing each data type now. Remainder don't know or have no plans. Based on 370 respondents.*

USER STORY CLEVELAND CLINIC BUILDS OUT ITS BIG DATA PROGRAM USING HADOOP AND DATA APPLIANCES

The Cleveland Clinic is consistently ranked as one of the top hospitals in the United States. A multispecialty academic medical center, it integrates clinical and hospital care with research and education. The clinic is also a data-rich, data-driven organization. According to Eric Hixson, senior program administrator for BI, the demand for analytics tools and platforms is growing at Cleveland Clinic. His organization has been pushing out dashboards and surfacing data to end users via data marts and guided data discovery. There is a large ecosystem of those kinds of assets.

The scope and scale of the underlying source data is growing, too. This includes electronic medical records (EMRs) as well as log data and data from sensors and machines. In order to utilize this data more effectively, Cleveland Clinic is enhancing its enterprise information infrastructure in a robust way through a multipronged strategy. "We are rolling out a high-performance data appliance for our warehouse and also utilizing Hadoop," Hixson explained. "On top of that, we're implementing an analytics infrastructure that includes predictive analytics, more modern machine learning, and AI capabilities so that the data appliances aren't just a new data warehouse—they are analytics enabling." This infrastructure includes both in-memory and in-database capabilities. Hixson's organization is also providing more robust data integration capabilities, including authentic metadata management and data quality, and data governance is a big part of the picture as well. "We want our end users to be able to use BI and statistical analysis tools to go after certified data sets so that their focus is on the analysis rather than guessing about the data," Hixson said.

> The main priority is to provide the right platform and the right data for those across the organization to address issues of patient care and risk. The clinic is looking at a distributed self-service model that includes a core, hybrid team of business analysts, data scientists, statisticians, and developers. The philosophy is that the infrastructure will support analysis using a range of tools that that the end user—whether administrator, scientist, or support staffer—is familiar with. The clinic is also looking to provide those who have the know-how with expanded access to data and tools. The key, according to Hixson, is that "the environment facilitates getting to the decision. We are looking to provide high-quality data in a performant-accessible environment."

Plans for Analytics

As mentioned, organizations are using big data and data science in numerous ways. These can involve using tools and techniques that are already in the organization as well as newer technologies such as NLP and machine learning techniques. In fact, NLP is used in text analytics tools today, while machine learning is often used for predictive analytics.

We asked respondents what kinds of analytics they are using today and what they plan to use in the future for big data (Figure 4).

Query and visual analytics are the most common use cases. Not surprisingly, query (and reporting) tops the list with 76% of organizations using it today. These organizations are likely making use of these tools against their increasing amounts of data for traditional reporting and analysis. Visual analytics was next with 53% using it today for big data analytics. Visual analytics enables users to do more on their own to analyze their data and answer their own business questions. Vendors are providing tools that allow users to visualize ever-increasing amounts and types of data.

Forty percent of respondents are using predictive analytics with big data now.

Predictive analytics is also in the top three. In addition to query and visual analytics, respondents also cited the use of predictive analytics with big data. This is a popular use case for data in general and where we see the market moving. In fact, in TDWI surveys, we typically see about 30%–40% of respondents stating that they use predictive analytics, with more having plans to do so. In this survey, 40% of respondents claim to be using predictive analytics for big data; another 48% state that they will be doing so in the next two years. Organizations use predictive analytics for numerous purposes including customer churn, fraud, customer targeting, patient readmission, and business process optimization. Predictive analytics requires a different way of thinking about data than descriptive analytics such as reporting or visual analytics. In predictive analytics, the mindset is about the probability of what might happen as opposed to describing what happened in the past. This requires skills discussed later in this report.

Respondents also have plans for geospatial and text analytics. In terms of analyzing multistructured data sources, respondents are also currently performing geospatial analytics (30%) and text mining (23%). As with predictive analytics, more plan to do so in the next few years.

Over 20% of respondents are performing text analytics today.

Geospatial data can be analyzed on its own using visual data discovery with layered maps. However, organizations also marry geocoded location data with other data for more advanced analytics. For example, risk models can be made more sophisticated by incorporating location-based information. Insurance companies might use predictive models to calculate the loss for a group of policies related to a possible weather event. Location-enriched data might include location features and characteristics such as soil type, proximity to rivers, home characteristics, and weather data such as rainfall amounts, storm intensity, or historical river floods. This location-based information can be used as part of the model to predict where the probability of payout will be high. The insurance company can then adjust premiums accordingly. Geospatial analytics is evolving, too. For instance, amusement parks offer bands that contain an RFID chip and a radio like those in a 2.4-GHz cordless phone. The band connects the wearer to sensors within the park and provides information about lines at rides. Nevertheless, this data is also analyzed to determine how customers move around the park, likes and dislikes, and so on. Retailers are deploying beacon technology to capture customers' movements around stores and offer them real-time promotions. The list goes on.

Likewise, organizations are making use of text data, often using NLP technology. A popular use case for text mining is voice of the customer. Here customer opinions, likes, and dislikes are analyzed from text data sources such as emails, call center notes, and social media data. Text analytics techniques are used to extract entities, themes, and sentiments from the data. This data is analyzed separately or, where possible, combined with other data about a customer to understand and take action on what customers are saying. Text analytics is also being used in other areas, such as in mining claims notes for fraud and patient notes to understand population health.

Streaming data and IoT analytics are lower on the list. Both streaming analytics and IoT analytics ranked lower on the list with 12% and 10%, respectively. Even so, if respondents stick to their plans, these numbers will be much higher in the next few years (an additional 35% and 30%, respectively). Streaming data is data in motion—that is, data that arrives continuously as a sequence of instances. This data comes from sensors, IoT devices, social media feeds, traffic feeds, and much more. Ingesting and analyzing continuous data streams enables a variety of organizational capabilities—including generating alerts for areas that need attention—using one or multiple streams as well as operational/situational intelligence to provide insight and visibility into an organization's business activities in real time. Some organizations analyze historical data using machine learning techniques, generate models, and then embed those models into streaming data for action.

What are your plans for the big data analytics listed below?

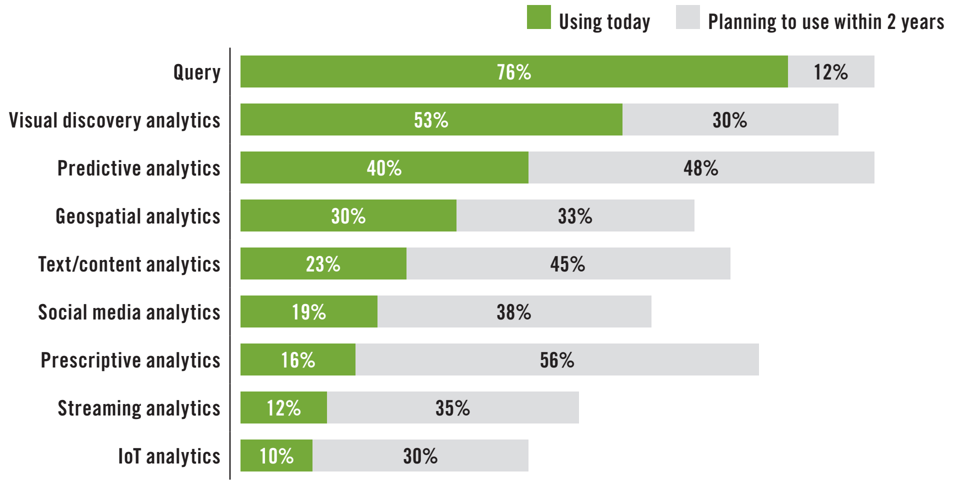


Figure 4. *Ordered by percentage using each data type today. Remainder don't know or NA. Based on 327 respondents.*

Data Management Systems Used for Big Data

Big data often involves new data platforms to support it. This includes MPP databases that can read many pieces of data across many processing units as well as newer platforms, such as Hadoop and NoSQL databases, to handle large volumes of structured, semistructured, and unstructured data. It might also include managing some of this data in the cloud. More often organizations are choosing to analyze data where it is created—increasingly in the cloud.

We asked respondents to tell us what platforms they are using now and planning to use to manage big data. Not surprisingly, 81% of respondents are utilizing their data warehouse to manage big data. This data warehouse, however, is evolving to support new technology and business requirements (Figure 5).[3] Other platforms include:

Organizations are using content management systems to manage big data, too.

Relational database on MPP and enterprise content management systems. Massively parallel processing refers to the use of a large number of processors to perform a set of coordinated computations simultaneously, in parallel. Data in MPP databases is partitioned across many nodes or servers. This can be quite useful for speed in big data environments where large volumes of data are involved. In this survey, 51% of respondents are using relational databases on MPP. Because big data is often unstructured data, it makes sense that many respondents are using some sort of enterprise content management system (47%) or document database (30%) as well.

Hadoop is being used by over 50% of respondents managing more than 10 TB of data.

Newer platforms emerge for on-premises use. Many respondents are looking to newer technology to help them manage their big data. For instance, 30% of all respondents reported using Hadoop on premises today. That number rises to just over 50% when filtering for only those who manage more than 10 TB of data. An additional 22% are planning to use it in the next few years. This is not surprising given that Hadoop enables computational analytics with massive, diverse data sets. The Hadoop family is available as open source from the Apache Software Foundation (www.apache.org), as well as from a number of software vendors that offer Hadoop distributions (or "distros") that package the Hadoop family, sometimes with additional tools and features for administration, security, and other value-added features.

Likewise, organizations are looking to newer kinds of databases such as NoSQL (23% currently), columnar databases (27%), and data appliances (25%) to manage big data.

The cloud is experiencing growth for data management. Respondents cited cloud options for big data management as well. As discussed earlier, it can make sense to manage data in the cloud that was created in the cloud, which is the concept of data gravity. This data can take many forms and might include social media data, sensor data, or data from cloud-based applications. Of course, cloud often becomes relevant in organizations precisely *because* of strains on data management. For instance, data warehouses designed for the cloud are being used by 28% of respondents currently. That number will double over the next few years if respondents stick to their plans. Likewise, Hadoop in the cloud is being used by 16% of respondents now, with another 28% planning to use it in the next few years. Respondents are also interested in options for data appliances and content management systems in the cloud.

[3] For more information on how the data warehouse is evolving, see *TDWI Best Practices Report: Data Warehouse Modernization in the Age of Big Data Analytics* (2016), online at www.tdwi.org/bpreports.

What kind of data management platforms are you using for big data now? Within two years from now?

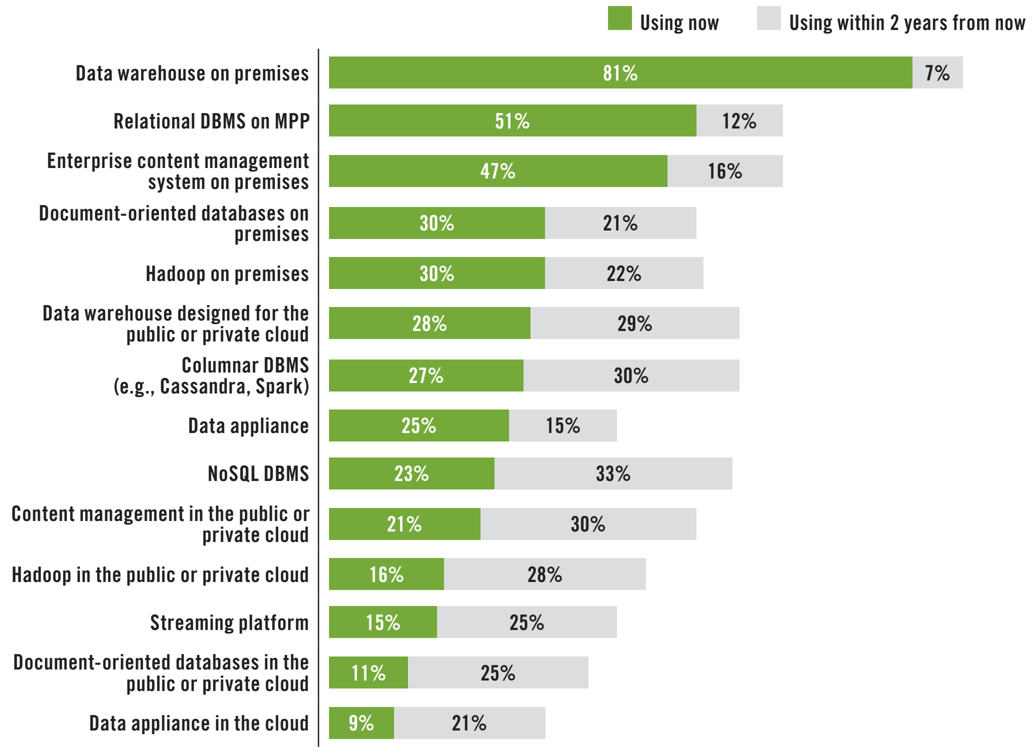


Figure 5. *Ordered by percentage using each data platform now. Remainder don't know or have no plans. Based on 338 respondents.*

Many respondents are lukewarm about their data management strategies.

Are respondents satisfied with their big data management strategies? We asked respondents to rate their level of satisfaction with their data management strategy on a scale from 1 to 5, where 1 is not at all satisfied and 5 is completely satisfied. As Figure 6 illustrates, many respondents are lukewarm or not satisfied with what they have in place. Why? In general, it seems that respondents felt that the strategy was not consistent across the organization. Many felt that there was fragmentation in how data was managed or that it was managed in silos. Others felt that big data and data management wasn't considered important enough by their organization. Some remarked that their strategy was sufficient for now, but it wouldn't scale. Funding was also an issue, as was data access. In order for data science to be successful, data management needs to be solid. Sometimes, in their zeal to get moving on the analytics side, organizations forget that data management is key to analytics success. Exploration on new kinds of nonvetted data is fine, but if that data is going to be put into production or used for decision making it needs to be managed properly, which includes making sure it is of good quality and comes from a reliable source.

Interestingly, respondents were likely to be more satisfied with their data management strategy as they built up a big data program and started using the technology and tools available, such as Hadoop, Spark, and others. These organizations have a strategy and the skills to make these new tools work, which seems to be paying off—at least in terms of satisfaction.

Using a scale from 1 to 5 where 1 means "not at all satisfied" and 5 means "completely satisfied," how satisfied are you with your current data management strategy?

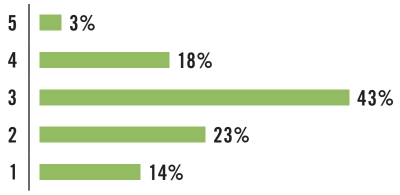


Figure 6. *Based on 338 respondents.*

USER STORY BOTH BUSINESS AND IT INVOLVEMENT ENHANCE THE DATA LAKE AT DELL EMC

In addition to the technologies cited above, the notion of a data lake has exploded into the market. A data lake is a repository for structured and unstructured data in a centralized environment that is used for analysis. Dell EMC, a leader in storage, converged systems, and the cloud, has had a data lake in production for about two years. According to Stephen Gatchell, Chief Data Officer, Engineering, Analytics & Data Lake, the company decided to move to a data lake environment for several reasons. First, its existing BI infrastructure felt too locked down and structured. Users (initially from sales and marketing and then other departments) wanted to analyze unstructured data such as social media data, customer requests, and log files. That was difficult in the current BI environment. Additionally, some existing systems were approaching obsolescence.

The team discovered that the data lake shouldn't simply be a dumping ground for data. Rather, Gatchell maintains that the most success occurs when "the focus isn't just on the data but what the business is using the data for. It is important to populate the data lake for use with business use cases." In this instance, the use cases necessitated both structured and unstructured data. Ultimately, there were two layers in the data lake. One is IT managed and contains scrubbed data that the company refers to as "hub data," which includes (but is not limited to) BI-related data for reporting. The other layer consists of what the company calls "innovation spaces," in which individuals own their own parts of the lake and can import and export data (including hub data) in a self-service environment.

As use of the data lake grew, the team realized that a key aspect of the data lake is how information is managed—and that includes data governance. According to Gatchell, "The business owns data governance because the goal of the overall effort is to improve total customer experience. The partnership with IT is very tight. We have an executive committee that crosses the business/IT line with joint collaboration—including budget discussions." In fact, the business/IT relationship is a key lesson learned in deploying a data lake. Gatchell believes that "most companies talk about getting data into a database but forget about business processes, business metadata, and data ownership. If someone wants to ingest new data, then they should own it." In this case, that involves the business. Gatchell sees a future where business becomes more technical and IT moves more toward the business—and perhaps the two won't be separated. For now, he knows that it doesn't make sense for IT to spin up environments that the business doesn't want or need.

Cloud Data Warehouses

As mentioned above, the cloud is playing an important role in big data management. Some organizations are moving their data warehouse to the cloud but may augment their existing data warehouse with additional platforms and tools such as Hadoop, MPP databases, appliances, data warehouses built for the cloud, or other kinds of data management platforms. The cloud then becomes part of a modern data ecosystem for big data and analytics.

We asked respondents about important features for cloud data warehouses. By far, the most important feature was security—cited as the top feature by 36% of respondents in a question specifically asking about the top feature (not shown). We've seen this in previous research related to analytics and the cloud as well.[4] Whether a real or perceived threat, security is always top of mind with respondents. This is not a bad thing because ultimately the organization is responsible for cloud security. Therefore, it makes sense to sit down with your cloud provider and ask questions about security practices, certifications, and controls.

Security tops the list of important features for cloud data warehouses.

We also asked respondents about other important features for cloud data warehouses. The answers (all not shown) reflect the fact that organizations are looking for scalability (26%), easy integration (26%), and speed in terms of loading (24%). Scalability is a top driver for the cloud in general. In order to use the cloud effectively, organizations want to be able to integrate cloud data sources with data on premises or in other cloud locations. They want to be able to load data quickly in the cloud. Organizations are also interested in providing cloud data access to analysts for use in their own self-service activities (30%).

Important Technologies for Big Data and Data Science in the Coming Year

We have been talking about big data and big data technologies and advanced analytics, but as mentioned above, data science includes an emerging set of tools and technologies such as Spark, R, and others. We asked respondents what top three technologies they felt would be important for big data and data science in the next year. The results are illustrated in Figure 7.

Hadoop, data warehouses, and open source R are rated as important technologies for big data in the coming year.

Not surprisingly, Hadoop tops the list with 41% of respondents choosing it as part of their top three. The data warehouse was not far behind with 35% of respondents citing it as one of the top technologies for big data. This is an important point: the data warehouse is not going away. It may be supplemented with other technologies, but organizations are not getting rid of the technology anytime soon. Other important technologies include:

Statistical open source package R. In our survey, 38% of respondents felt R was very important for big data analytics. As described earlier, R is a popular open source language and statistical environment. Many universities use R as a learning platform for analytics. It contains numerous statistical algorithms and methods. Many organizations experiment with analytics projects using R. Although the jury is still out about deploying R into production due to some performance issues, there is no doubt it is an important tool for analyzing big data. In fact, many vendors are providing support for R in their products now so that R programming features can be used in their software packages or against their data management systems.

Spark. Spark is also receiving quite a lot of market attention. As mentioned, it is an open source in-memory processing engine known for its speed. It also has a sophisticated analytics library and supports streaming (Spark streams). Twenty-two percent of respondents felt that Spark would be very important in the coming year. Vendors are also embracing and supporting Spark for big data processing and application development. Some provide connectors to Spark so their data management systems can be used as a data source for Spark.

More than 20% of respondents felt that Spark would be an important open source technology in 2017.

[4] For more information on security in the cloud, see *TDWI Best Practices Report: BI, Analytics, and the Cloud* (2016), online at www.tdwi.org/bpreports.

Python. Although not as popular as R in our survey, about 17% of respondents felt that Python would be an important technology in the next year. As noted, Python is an easy-to-use general-purpose programming language with a number of libraries that can be used to build analytics applications.

The cloud. The cloud has also been mentioned as the go-to platform for big data. We've already seen that the cloud is becoming important in big data management and analytics. More often, data is being created in the cloud, and more often, data scientists want to analyze it there. They may want to explore the data and look for important attributes, then bring that data on premises for further analysis (this is the concept of attribute reduction—that is, if the data scientist starts with 1,000 attributes, he or she may be able to reduce this to a smaller number before bringing it on premises). For some applications, such as IoT analytics, organizations may want to analyze data in the cloud to avoid the cost of moving so much data on premises, or the data may be changing so fast there may be no need to move it on premises. In some cases, data is even being analyzed at edge nodes in the network.

Often data is being created in the cloud and data scientists want to analyze it there.

In this survey, 17% of respondents felt that the hybrid cloud would be important for big data in the coming year. The hybrid cloud is a computing environment that includes the use of public and private clouds, often with one or more touchpoints between them. It might include some sort of integration process to enable cloud and on-premises environments to work together. In a recent TDWI Best Practices Report, 43% of respondents thought that their architecture would evolve to a hybrid ecosystem approach where the cloud might form one component of a bigger architectural strategy.[5]

Data virtualization. This process provides an integrated view of data from disparate sources into a single "virtual layer" without replicating the data. In other words, it provides an abstraction layer that makes it easier to integrate data from multiple big data sources/stores. In our survey, 21% of respondents felt this would be important in the coming year.

What technologies/tools/languages do you believe will be very important for big data in the next year? Choose up to three.

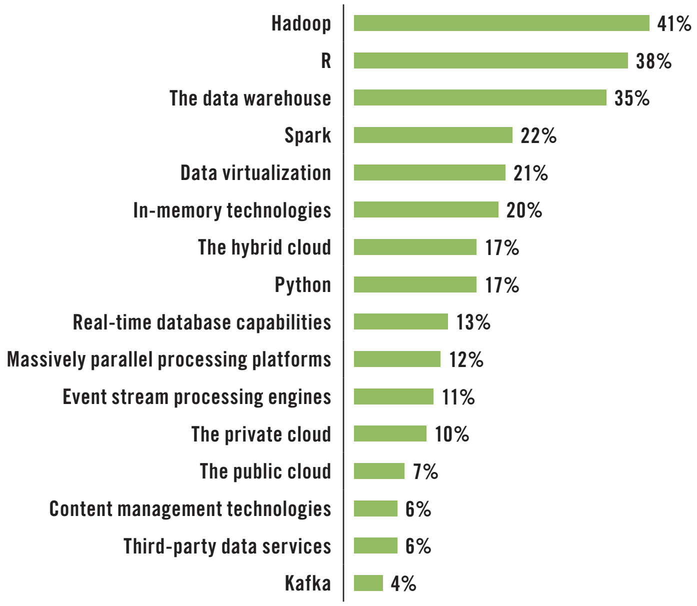


Figure 7. *Based on 327 respondents.*

[5] For more information on the cloud, please see *TDWI Best Practices Report: BI, Analytics, and the Cloud* (2016), online at www.tdwi.org/bpreports.

USER STORY BUILDING A DEDICATED HEALTH ANALYTICS FUNCTION AT REGION OF PEEL

The Region of Peel serves approximately 1.4 million residents and 135,200 businesses in Brampton, Caledon, and Mississauga, Canada. The Region's Health Services Department provides services such as Public Health, Paramedic Services, and Long Term Care. According to Jamie Barnes, Manager of Health Analytics at the Region of Peel, analytics has always been performed in specific program areas to manage service delivery and report on transactional data. However, while Health Services had been generating a lot of data and analyzing it for particular projects, they weren't doing it consistently and with a focus on trying to maximize value from data across the whole department by mining it in more detail.

The Commissioner of Health Services and other administrators thought a dedicated analytics group that would work together with other areas of the organization that play a role in managing the corporation's data was a step in the right direction. While the group is just getting off the ground, its goal is to provide broad and deep analytics skills for Health Services. The group works on a project-by-project basis and meets frequently with managers across the department to determine what projects will add value. To show quick success, the group is working on low-hanging-fruit projects that currently have centered on understanding paramedic activity and gaining insight into some seniors' services.

Barnes said the launch has had its issues. There has been some confusion about what his group does and how it works with other groups in Health. He said they are careful to "communicate that we're not replacing functions such as program-based analytics groups." He is also in the process of trying to build out a data science team and realizes that there is no one "unicorn." Instead, he is looking for people with skills in computer science, mathematics, spatial analysis, and quantitative evaluation. Barnes believes they will "build these four pillars and then draw on certain aspects of those depending on what we need on a project-by-project basis."

Barnes adheres to several philosophies. The first is the group won't be restricted by tools. "It shouldn't be about the tools. It is hard enough to find people with the right knowledge let alone the right knowledge and the particular set of tools you've arbitrarily chosen. They should use whatever packages allow them to do their best work. If R is the tool they know, they should use that. If they can work miracles with SPSS or Matlab or another combination of open source packages, then that should be fine too," Barnes said. He also is a supporter of the "fail fast" notion. According to Barnes, it's a numbers game. "If I know 1 out of 10 projects will succeed, then I need to get through them to maximize opportunities, to maximize the chances of discovering that one good idea. Let a thousand flowers bloom." Part of this requires a prototyping and experimentation mindset within the context of the business. Barnes is a big believer that analyzing data requires knowledge of the business and understanding the context. These principles and others will help guide the group moving forward.

Open Source Technologies in Data Science and Big Data Analytics

We've seen that many respondents believe open source technologies such as R, Hadoop, and Spark will be important for big data and data science. In order to dig into this in a bit more detail, we asked respondents about their views on open source. The open source model is a collaborative development model where code is freely available and the copyright holder has the rights to study, change, or distribute the code. Open source has become quite popular, especially for big data and data science, because it is a low-cost source community for innovation, which appeals to many data scientists and analytics application developers.

R, Hadoop, and Spark are important technologies for big data and data science.

Open Source in Use Now

We asked respondents about their use of Apache open source software (Figure 8). Not surprisingly, Hadoop and Spark rank in the top three. Other open source technologies in use include:

- **MapReduce.** Developed at Google, MapReduce is a framework that provides processing scalability across huge numbers of nodes in Hadoop. In MapReduce, the data is distributed over the cluster and processed. Spark is in-memory; MapReduce is not. Spark is often used for streaming data and MapReduce is used for batch. Although many feel it is on its way out, 24% of respondents are using MapReduce now and 22% expect to use it two years from now.
- **Hive.** Originally developed at Facebook, Hive is a data warehouse infrastructure built on top of Hadoop. It provides interactive query capability over data in Hadoop. Twenty-three percent of respondents are using Hive now and an additional 22% plan to use it in the next few years.
- **HBase and Cassandra.** Less than 15% of respondents are using open source distributed databases such as HBase or Cassandra. That number might double if users stick to their plans.

Is your organization using or planning to use any of the Apache open source software listed below for its big data efforts?

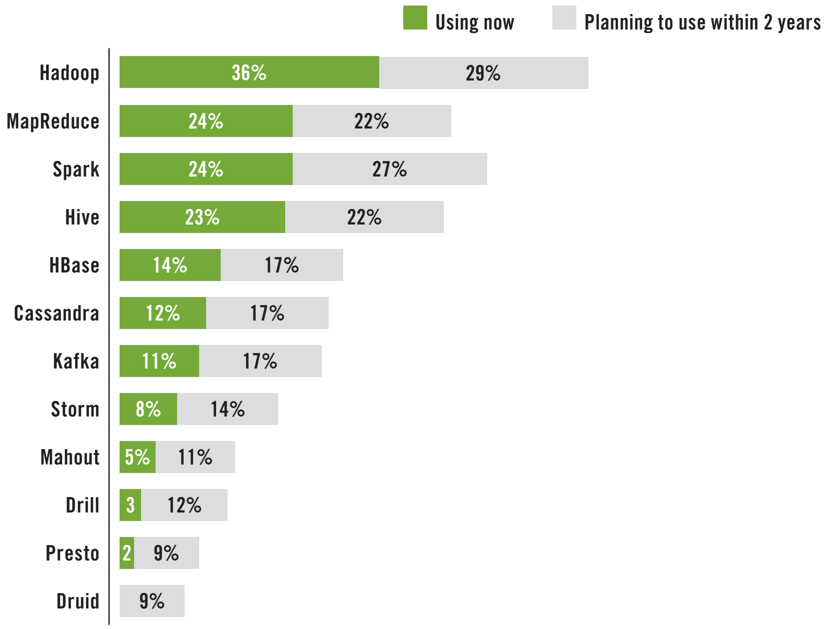


Figure 8. *Ordered by percentage using each software now. Remainder don't know or have no plans. Based on 338 respondents.*

Opinions of Open Source for Data Science

We also asked respondents about their opinion of open source technologies and whether they can be used in production for big data and data science (Figure 9). The pro-open-source community likes the fact that it is innovative, flexible, low cost, and interoperable. For analytics, supporters like that in addition to providing powerful tools for big data, the more popular technologies have large communities behind them, which makes it easy to find answers to problems. Going with an open source platform allows a company to be independent from a vendor's proprietary software stack. It also allows them to be part of a community.

The majority of respondents find open source most useful with added functionality for reliability and scalability.

Those who are not fans of open source often cite that it can be difficult to use, is not supported, isn't reliable or secure, and doesn't scale. In particular, in analytics, open source has the reputation that it is good for experimentation but not necessarily for production, because it doesn't include features or functionality needed for production such as security or good performance.

In our survey, respondents overwhelmingly seem to like open source options. Forty-seven percent stated that open source helps build skills and leverages others' skills. They believe it can be deployed in production. That said, 60% of respondents find open source with added innovations that make it reliable and scalable (i.e., commercialized open source) to be the most useful (not shown).

Which of the following most closely resembles your opinion with regard to open source and big data/data science efforts?

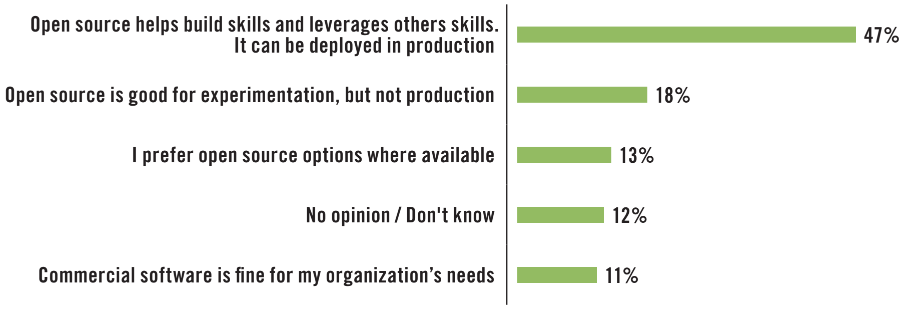


Figure 9. *Based on 338 respondents.*

Clearly, however, there is not one strategy for delivering on big data projects, nor is there one clear path to value (discussed later in the report). We asked respondents how they believe their organization will deliver on big data analytics (Figure 10). Respondents could select more than one option. Although respondents like open source and 54% could see delivering on big data analytics using an open source analytics tool like R, 56% said that some big data projects would be delivered using commercial software on premises or in a private cloud. Other options included public cloud SaaS tools (34%) as well as custom developed apps (44%).

How is or do you believe your organization will deliver on your big data analytics projects? Please select all that apply.

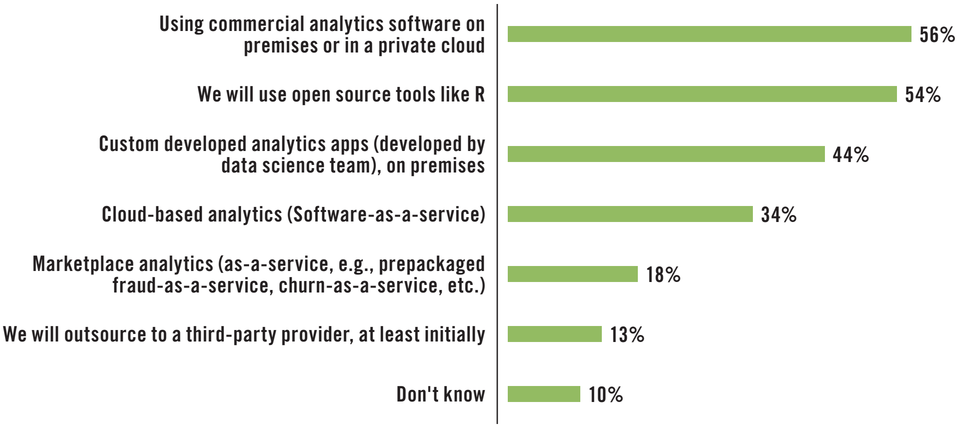


***Figure 10.** Based on 327 respondents.*

Organizational Strategies for Data Science Success

We have been discussing the data, platforms, and analytics that organizations are starting to utilize as part of their big data programs. Some of these incorporate traditional data and methods; others are newer. As organizations embark on their data science initiatives, they often struggle *more* with the organizational aspects of big data and data science than with the technology aspects. To find out more, we asked several questions about leadership strategies, organizational structures, and how respondents plan to build data skills in their organization.

Leadership Strategies for Big Data and Data Science

Nearly 30% of respondents have a VP of analytics leading their big data efforts.

Putting any kind of new program together often requires executives who can set the tone and vision and drive the effort. They can also provide much-needed funding support. The same is true for data science initiatives. Over the past few years, there has been a lot of market talk about the CAO (chief analytics officer) and the CDO (chief data officer). The CAO is responsible for the results of analytics projects. This person is often the visionary evangelist who is also adept at change management. Likewise, the CDO focuses on big-data-related activities, including the infrastructure to support big data, and sets the vision and execution of big data management.

We asked respondents who already have a big data program in place or plan to have one in the next six months (142 respondents) who in their organization was leading their data science effort. Twenty-eight percent said that the VP of analytics leads their data science efforts. Seventeen percent said the CIO leads the efforts. About 5% each (15%) cited the “chief” positions—the CDO, CAO, or chief data scientist (all results not shown).

Regardless of who the leader is, a leader is important to the success of big data and analytics initiatives. Someone has to own the effort in order for it to mature and grow. As the old saying goes, “If everyone is responsible, then no one is responsible.” We have seen that organizations can put these programs in place from the bottom up. However, it generally takes more time to do it this way than when there is an executive sponsor. Of course, it needs to be the right executive—someone who understands or at least appreciates and wants to learn about these technologies. The sponsor will also be important for setting reasonable expectations about what can be delivered.

USER STORY MAKING THE FUNDING CASE FOR BIG DATA ANALYTICS IN BIG PHARMA

Although organizations are concerned about data and analytics skills, the overall cost of big data analytics programs is consistently ranked highly by respondents as a primary barrier to implementation (cited by 28% of this survey's respondents). TDWI recently spoke to the senior director of big data analytics at a major pharmaceutical company about her group's approach to the funding challenges it faces implementing big data in decision making.

According to the senior director, "Funding and resourcing are some of our biggest constraints." To overcome this challenge, the group currently obtains funding from more than one department—in this case, corporate and R&D. "We were in agreement to split the cost between the corporate headquarters and R&D," she said. This kind of split cost structure is one approach taken by companies in the process of implementing big data solutions.

Here are some best practices this company applied to alleviate the funding issue:

Start small and build a business case. "I've positioned projects as proof of value and we start small." She underscored the importance of asking, in a universal sense, "What is the business question we are trying to answer?" Instead of getting bogged down with the minutia of process, they've gotten traction by keeping people attuned to bottom-line impacts.

Rely on evidence-based support from those more accepting of the practice. "The challenge that I've had in talking to other areas [outside of the scientific practitioners] is that they want a guarantee that something will come out of it that's 100% applicable right away. As we know in data science, some of it is just exploratory." Because some departments within the organization might not readily see the benefits of big data, it is important to seek out people who do and disseminate results. "The scientists actually understand it more—that experimentation is trial and error—and are willing to take more risks in terms of failure." Using the resulting discoveries from big data analytics from those who are willing and able to undertake the practice now is a powerful means of making a sound business case for those who might be more reticent.

Talent-Building Strategies

Respondents are using a range of approaches for finding and developing talent.

One of the biggest challenges that organizations face in terms of implementing big data and data science programs is finding the right people with the right skills. In fact, when we asked respondents what the biggest organizational challenge is or will be for big data and data science, skills for analytics topped the list, with 40% citing this as a top challenge (Figure 12).

The "true" data scientists require numerous skills—from knowledge of statistics and advanced math to computer science and development. They also need to be critical thinkers who know how to communicate and who understand the business. These are people who understand the technologies discussed in this best practices report—from machine learning to NLP, and from programming in R to running in Spark. This is a tall order.

How do organizations find the right skills for their big data and data science efforts? The respondents to this survey use a multifaceted approach for building a talent base (Figure 11):

More than half of respondents stated they will grow big data and data science skills internally.

Build talent from within. Fifty-one percent of respondents said they will grow big data and data science skills internally by enhancing existing business analyst skills. This is the idea behind the emergence of the citizen data scientist. Because the data scientist is a rare commodity, organizations feel they can train members of their existing team to become more technical. They use a range of approaches from self-teaching to online or onsite learning. Some send their employees to boot camps or other learning venues. Those employees who become quite interested are sometimes able to pursue an advanced degree, if the organization can afford it.

Another factor contributing to the rise of the citizen data scientist is that vendors are making their tools easier to use. For example, some vendors provide tools where the user simply specifies target or outcome variables of interest for a predictive model. The system then determines the best model given the attributes. Some tools even explain the results!

This notion of the citizen data scientist is both good and bad. On the one hand, if business analysts can expand their skill set that is a good thing. If a tool is easy to use, then so much the better. Nonetheless, it is not good for someone to use highly advanced analytics tools without any training. If they are going to try, the organization needs to put controls in place before a model created this way goes into production. For instance, some organizations use data scientists working with business analysts as the control point. The data scientist needs to OK the model before it is put into production.

Hiring data scientists from outside the organization. Thirty-six percent of respondents are already hiring or planning to hire data scientists from outside their organizations. As stated previously, the data scientist is a rare breed. One approach organizations use is to hire a few data scientists and then supplement them with business analysts/citizen data scientists. Other organizations outsource the data science function. In this survey, 13% of respondents are outsourcing data science talent for the time being, often done to build a proof of concept. Although it can take time, there can be a transfer of knowledge between the consultant and the organization.

Building data science teams. Because there are so many skills that a data scientist needs, it often makes sense to use a team approach to data science. In this survey, about 20% of respondents stated that they might use this method to build up their talent pool.

What are the top two strategies your organization is using/planning to use to hire data scientists? Please select two responses.

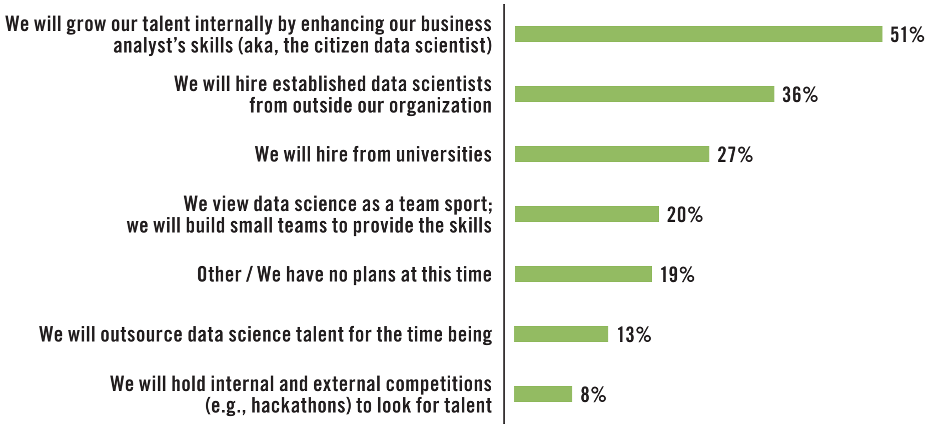


Figure 11. Based on 327 respondents.

EXPERT OPINION KNOWLEDGE TRANSFER IN OUTSOURCED DATA SCIENCE PROJECTS

Skills are extremely important in big data and data science projects. Some organizations will outsource these projects in the initial stages because they don't have the skills. The thought is to then to do some sort of knowledge transfer to bring the projects back in-house.

Dean Abbott, president of Abbott Analytics, explained that although he likes "the idea of knowledge transfer, if the problem is anything other than a vanilla data-mining problem, it can take months before the company 'gets it' and can be productive. Knowledge transfer is tricky because *what* knowledge to transfer is critical. It can take considerable time on the consultant's part to figure out how to communicate their findings in a way the stakeholder can really understand, and unless the stakeholder has a good understanding of analytics, it may just go over their head. The first step for an organization that wants to outsource is to have the managers of the data-science outsourced team learn data science. I fear that unless that happens, the projects will fail—unless the consultants are able to pick up that slack."

Why is this the case? "While knowledge transfer makes sense in principle (i.e., that the internal staff could learn and take over doing the predictive modeling), I never saw them succeed, except with the completely vanilla modeling projects," said Abbott. "Sure, they got better at it, but most often the internal staff never progressed to the point where they could completely take over. There were some notable exceptions, but they were just that: exceptions."

"One thing I think they learned is that data science/predictive analytics/advanced analytics is hard to do well, and it takes more than raw skills to be successful. In other words, it takes more than learning how to use a software package, building a linear or logistic regression model, and deploying code to have a successful analytics practice. I've come to believe that there is a mindset that has to be there as well, what I often call the '*Freakonomics* mindset.' This includes a combination of intellectual curiosity, creativity, and attention to detail (to find those telltale signs that the data is bad). Many folks I worked with learned the science but never mastered the art."

Centers of Excellence

Many organizations build out a center of excellence (CoE), which consists of a cross-functional group that provides leadership in big data and/or data science. In addition to building and deploying analytics, CoE teams are often responsible for training and disseminating best practices. In this survey (not shown), about a third of respondents had a CoE in place. An additional 26% were planning to deploy one in the next year.

At least a third of respondents already had some sort of a CoE in place.

Some CoEs are companywide and may have teams within the center that serve different business areas. Business units commission the work. Other organizations distribute analytics experts throughout the organization into the lines of business. Some use a hybrid approach. The best method is still under debate; however, if the CoE is decentralized then there needs to be a process in place for sharing best practices. In this survey (all not shown), about 18% of those with a CoE used a centralized approach where the analytics professionals are located in the CoE. Another 20% used a distributed approach where analytics professionals are embedded in the business units and report to that business unit. Less than 10% used a distributed approach where the analytics professionals are embedded in the business unit but report to a central person, such as the CAO.

Delivery Strategies

Finally, who is building models with big data? In previous studies, TDWI has seen a shift to the business analyst as a builder of more advanced analytics models. As mentioned, business analysts are being trained in some organizations to build their skills and to supplement data scientists who are brought into the organization. Others are using more easy-to-use tools for this kind of analysis.

In this survey, the business analyst and the data scientist/statistician are both analyzing big data using more advanced analytics. Sixty percent of respondents cited the data scientist/statistician as analyzing data using advanced analytics such as predictive analytics, text analytics, and so on. Fifty-four percent of respondents cited business analysts as doing this kind of analysis (all not shown).

Challenges Ahead

Despite the benefits of big data in terms of better insight, understanding, and competitiveness, respondents noted barriers as well. Some of these are related to organizational issues; some are technology focused. We asked two separate questions about challenges (Figures 12 and 13), which are discussed in more detail here.

Organizational Challenges

Skills (40%) was cited as a top organizational barrier to big data deployment.

Respondents were asked to pick the top three organizational challenges they have faced or believe they will face when deploying big data. The top organizational barrier to data science and big data is skills related to analytics (40%). As described earlier, skills for data science are in short supply. Respondents spoke about “a lack of skilled analysts to leverage big data tools” as well as “a lack of in-house skilled personnel, which results in overreliance on contractors.” The other organizational challenges focused on data—getting access to it because of politics and other considerations (36%) as well as governing it (29%). For instance, some respondents talked about problems instituting data governance policies “across disparate groups and data sets.” Others talked about governing the citizen data scientist.

Interestingly, big data access (45%, not shown) and analytics skills (40%, not shown) are still the top two challenges cited by those organizations who claim to already have a big data program in place, but data management skills moves up to take the third spot (27%, not shown). Governance moves to fourth place.

What are/were/do you think will be the biggest organizational barriers to adoption of big data in your organization? Please select up to three responses.

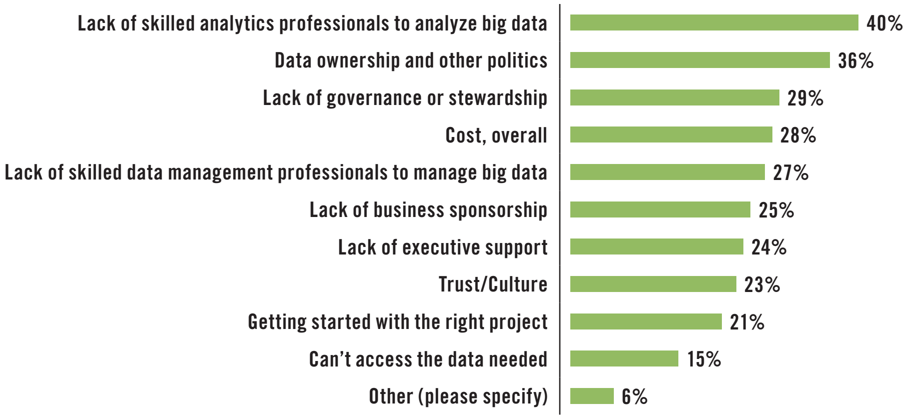


Figure 12. Based on 352 respondents.

Technology Challenges

Likewise, respondents were asked to pick the top three technology challenges they have faced or believe they will face when deploying big data. The top technology barriers to adoption of big data are lack of understanding of big data technologies (40%) followed by lack of enterprise data architecture (38%) and concern about big data security and privacy (36%). There is also concern about integrating big data with other data sources (34%). Respondents cited issues associated with understanding technology such as "too many new technologies and too much hype" as well as "fast-changing data architectures." Larger organizations often have issues with "[the fact that it's] too large and can be slow to adopt—it is hard to get past risk assessment and the security requirements are so high it becomes very costly." This speaks to the knowledge and skills barriers described previously, as well as some of the political and cultural issues that can slow down any analytics project.

Not understanding big data technologies was cited by 40% as a barrier to adoption.

What are/were/do you think will be the biggest technology barriers to adoption of big data in your organization? Please select up to three responses.

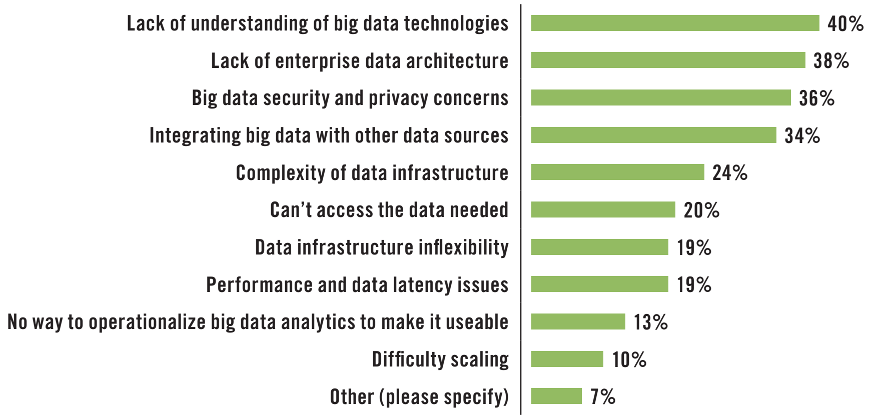


Figure 13. Based on 352 respondents.

Paths to Value

There is no single path to success for big data and analytics.

There is no single path to success for big data and analytics. As we have seen, there are challenges around skills, organizational models, and technologies. Organizations are utilizing different strategies for platforms, tools, and organizational issues to get big data projects off the ground and keep them advancing.

Overcoming Challenges

We asked respondents how they are overcoming some of these challenges. Their responses fell into several groups.

- **Education and training.** Respondents are planning to not only hire the right people but also train from within. Many are planning to "look for candidates and send motivated employees to training." Although in-house training and self-learning can be helpful, many organizations realize that it is important to send employees outside to obtain training and certifications by skilled instructors. In addition to training is education. Many respondents spoke about the need to educate their organization about big data and data science. Some spoke about education along with small proofs of concept to show value (see below). Others mentioned educating stakeholders and executives about the value of big data to build excitement and awareness.
- **Collaborate and communicate.** In addition to education, respondents also focused on the need to collaborate and communicate to build trust. Respondents spoke about the need to "communicate with the right people with the right attitude at the right level in the organization," as well as to "provide regular updates to the joint stakeholder group." Persistence, collaboration, and communication are key to moving big data projects forward.
- **Proof of concept (POC).** Many respondents talked about building small POCs and, in particular, "POC projects to show the difference." In other words, these POCs use real business problems or small use cases with definitive business value to prove out the benefit. These POCs are sometimes considered quick wins that show ROI as soon as possible.

Driving Measurable Value

In order to explore best practices further, we examined the characteristics of companies that are gaining value from big data and data analytics. We also wanted to see if those companies differed from companies that have not reported measurable value. We separated those who measured value (54 respondents), those who think they have gained value but couldn't measure it (144 respondents), and those who did not measure value (24 respondents) into separate groups. The rest either didn't know or stated the question was not applicable at this time. These are small numbers and should be considered preliminary results at best. Even so, some interesting results did emerge from looking at the data in this way.

- **Center of excellence (CoE).** A CoE is an organizational model that can provide substantive value. In this survey, those who measured top- or bottom-line impact were more likely to have a CoE in place than those who did not (e.g., those who think they gained value and those who measured no value).

- **Growing talent internally.** Many organizations are looking to grow their data science talent internally. In this survey, those who measured a top- or bottom-line impact were more likely not to grow their data science talent internally. This is not to say that business analysts can't be trained to do more sophisticated analysis. Still, it is important to find the right business analysts and to train them properly.
- **Advanced analytics.** Industry studies have indicated that advanced analytics drives value. We have seen this in TDWI research as well. In this survey, those who measured top- or bottom-line impact were more likely to utilize advanced analytics such as predictive analytics than those who did not (e.g., those who think they gained value and those who measured no value). This also seemed to be the case with some other advanced analytics, such as text or streaming analytics, and with open source R—although further analysis with more data is needed.
- **Disparate data sources.** Those who measured top- or bottom-line impact were more likely to be collecting/analyzing disparate data types such as text data or semistructured data than those who did not (e.g., those who think they gained value and those who measured no value).

There were also some weak correlations between open source technologies such as Hadoop or Spark and value, but it is probably still too early in market adoption to speculate on this. Likewise, having a chief officer of analytics, data science, or data in and of itself did not seem to drive value.

Of course, correlation does not imply causation and these numbers are low; the correlations are significant but weak. The organizations that have measured value may be more mature analytically—hiring data scientists and making use of them—and they may be expanding their data ecosystems. They also may be more deliberate about their data science and big data efforts than those who haven't measured value, and quantifying impact may be part of this process.

In sum, there are different paths to big data and data science success. Some organizations use commercial software; others use open source. Many use both and both provide value. Many organizations are seeing that the cloud may prove useful for big data and data science. This seems to be evolving toward a hybrid cloud model. Organizations are adopting technologies such as Hadoop and Spark, yet the data warehouse is not going away as it still provides value. Some organizations make use of both business analysts and data scientists/statisticians to build more advanced analytics. This might work if the correct controls are put in place and the business analyst is properly trained. The paths to big data value are wide and varied and depend on the organization and the business problem it is trying to solve, as well as where the organization is in its current stage of evolution.

A Sample of Relevant Vendor Platforms and Tools

Because the firms that sponsored this report are all good examples of vendors that offer technologies to support big data and data science, we present a brief look at the product portfolio of each. The sponsors form a representative sample of the vendor community, although their offerings illustrate different approaches.[6]

IBM

IBM provides a broad range of capabilities for big data and data science. These include IBM BigInsights for Apache Hadoop, IBM BigInsights on Cloud, and IBM Streams. The company believes that that the market is moving from products to platforms and that unlocking data at scale requires a different approach—a platform-based approach rather than a modeling approach. It recently launched the IBM Watson DataWorks platform: a cloud-based platform that provides data access and analytics capabilities for multiple data types. IBM is using cognitive capabilities organically as part of the platform. Project DataWorks is founded on being open with an open ecosystem that enables application development. The platform includes Apache Spark, IBM Watson Analytics, and the IBM Data Science Experience—a new development environment in the cloud for real-time, high-performance analytics. DataWorks is available on IBM's Bluemix cloud platform.

MapR

MapR provides a converged data platform that runs a variety of workloads and use cases in a single deployment. The platform integrates Hadoop and Spark with event streaming, real-time database capabilities, and enterprise storage to support big data applications. Organizations use the MapR Platform for a variety of use cases from pure data warehouse offload to machine learning to real-time Internet of Things analytics using streaming data. MapR platform services provide a range of data handling capabilities. It packages and supports a broad set of Apache open source projects. MapR tests and integrates open source ecosystem projects such as Hive, Pig, Apache HBase, and Mahout, among others, as well as partnering/integrating with commercial platforms such as SAS, HP, SAP, and Cisco.

OpenText

OpenText provides the OpenText™ Analytics Suite to support big data and big data analytics. The OpenText Analytics Suite is comprised of two integrated products that work in tandem to support big data: OpenText Big Data Analytics and OpenText Information Hub (iHub). iHub is a BI and analytics platform that enables organizations to design, deploy, and embed reports, visualizations, and interactive dashboards into any application and on any device.

Big Data Analytics extends the functionality of iHub to provide business users and analysts with self-service capabilities for data preparation and data exploration, as well as advanced and predictive analytics. Nontechnical users can access, blend, explore, and analyze their data without depending on IT or data experts, then share and socialize their insights. It is built on a columnar database that can manage disparate data sources. It includes built-in statistical techniques for a range of analytics including geospatial analytics, predictive analytics, and more. Both products are offered on premises or in the cloud.

[6] The vendors and products mentioned here are representative. The list is not intended to be comprehensive.

Snowflake Computing

Snowflake Computing provides a cloud-native data warehouse solution delivered as a service that utilizes an architecture built specifically for the cloud. Snowflake's multicluster, shared data architecture decouples data storage, query processing, and metadata management, making it possible to bring together data in a single location while independently scaling computing horsepower on the fly. The data warehouse can be scaled up or scaled down at any time without redistribution of data, read-only downtime, or delay. Snowflake supports big data and was designed for semistructured and structured data that can come from both discrete and streaming sources. The service allows users to load semistructured data without having to define a fixed schema and then query that data in combination with structured data using SQL. A recent addition is a multicluster warehouse capability, where a customer can predefine a minimum and maximum number of clusters and allow Snowflake to scale up and down that number automatically to support varying levels of concurrency with consistent performance.

Top 12 Best Practices for Big Data and Data Science

In closing, we summarize the report by listing the top 12 best practices for big data and data science along with a few comments about why each is important. Think of the best practices as recommendations that can guide your organization into successful implementations of big data and data science.

1. **Get your data in order.** The right data management strategy is important to big data and data science success. In the zeal to get started analyzing data, organizations often don't pay attention to that data. Yes, it is OK to experiment on raw data; a good data scientist usually explores the data before building models—particularly models that are put into production. However, we've seen that access to data is a challenge for those embarking on big data and data science, which might be due to politics or data integration issues. It is important to make sure the data is in order. That will ultimately include collaboration between different parts of the business as well as governance.

2. **Plan on a phased approach.** As mentioned above, many respondents cited the value of a proof of concept. That use case, when it succeeds, should be designed to provide a lot of value. Success begets success. Don't try to boil the ocean—there is too much data there. Plan and execute in phases.

3. **Get some training.** This can't be stressed enough. As cited above, a big challenge to big data and data science is knowledge. This is true for those looking to deploy new data management platforms as well as those planning to analyze big data. Even if tools are supposed to be easy to use, typically they are not. It is important to understand how technology works and how advanced algorithms operate. If you're using a machine learning algorithm, understand it first. Before deploying NLP, make sure you know how it works, as well as its strengths and weaknesses.

4. **Move past the data warehouse.** The data warehouse is not going anywhere anytime soon. Nevertheless, big data may necessitate moving beyond the data warehouse to platforms that can support multistructured data and iterative analytics. That said, don't be seduced by every new

big data platform. Before adopting one, be sure it can satisfy real-world requirements with the right performance and in a cost-effective manner.

5. **Use disparate data types.** Although structured data is still the mainstay of modelers and analysts, disparate data types can enrich a data set and provide lift to models. Think about incorporating new kinds of data, such as text data and geospatial data, into the mix. Depending on your use case and business needs, streaming data can also be quite valuable for situational awareness and improving operational efficiencies. Of course, you'll need the right tools for the right jobs.

6. **Use multiple analytics methods.** Organizations are starting to move beyond basic reporting and dashboards and that is a good thing. Analytics, such as predictive analytics, can provide real value. However, many organizations get hung up on predictive analytics as the goal for analytics. There are other kinds of analytics that can be used in conjunction with (or separately from) predictive analytics. These include text mining, geospatial analytics, and graph analytics. All can provide value and those who are most successful make use of multiple kinds of analyses.

7. **Consider a center of excellence.** As described earlier, a CoE can be a great way to make sure that the infrastructure and analytics you implement are coherent. CoEs can help your organization disseminate information, provide training, or maintain governance.

8. **Consider open source technologies.** Open source technologies can provide a cost-effective way to gain access to a large community of innovators. These technologies can be worth exploring, although they require a certain skill set.

9. **Consider the cloud as part of the data and analytics ecosystem.** Some organizations will not move their data or analytics to the cloud (especially the public cloud) because of security concerns, yet many cloud providers (especially the large ones) have better security than that found on a company's premises. Organizations that have moved to the cloud often reap the benefits of scalability, flexibility, and agility—especially for big data. It is worth exploring and asking questions of cloud providers about this option.

10. **Address cultural issues.** Change can be hard. Education is critical here, as is changing the mindset. Some people don't get it. Some have legitimate concerns. Some are concerned about their jobs. It will be important for those driving change to get executive support (someone who is the champion) and help to communicate their message.

11. **Plan for new architectures.** Big data will necessitate new platforms and new architectures. These evolving ecosystems might include the data warehouse, Hadoop, and other platforms, both on premises and in a public cloud. Data scientists, citizen data scientists, and others will need to access data from different sources. The architecture can become complex, but can be manageable if a plan is put in place. This means reworking architecture plans to determine how platforms will integrate and operate together.

12. **Take action on big data analytics.** What good is all of the insight gleaned from big data unless you take action on it? Insight from a PowerPoint presentation or a visualization tool is great, but making the analytics developed through a big data effort part of a business process is where real value will occur. Think about how you might operationalize or embed analytics into a process to help drive or automate action on analytics.

IBM
www.ibm.com

IBM has the most complete set of capabilities to enable today's hybrid data warehouse, spanning from on-premises appliances (PureData System for Analytics) to Hadoop solutions (BigInsights) to data warehouse (dashDB) deployments that address virtually every information need: structured, semistructured, or unstructured data, as well as hybrid deployments. To learn more visit: ibm.com/data-warehouse and ibm.com/Hadoop.

MapR
www.mapr.com

MapR enables organizations to create disruptive advantage and long-term value from their data with the industry's only converged data platform, which delivers distributed processing, real-time analytics, and enterprise-grade requirements across cloud and on-premise environments—while leveraging the significant ongoing development in open source technologies including Spark and Hadoop. Organizations with the most demanding production needs, including sub-second response for fraud prevention, secure and highly available data-driven insights for better healthcare, petabyte analysis for threat detection, and integrated operational and analytic processing for improved customer experiences, run on MapR. A majority of customers achieve payback in fewer than 12 months and realize greater than five times ROI. MapR ensures customer success through world-class professional services and free, on-demand training that over 50,000 developers, data analysts, and administrators have used to close the big data skills gap. Amazon, Cisco, Google, HPE, Microsoft, SAP, and Teradata are part of the worldwide MapR partner ecosystem. Investors include Future Fund, Google Capital, Lightspeed Venture Partners, Mayfield Fund, NEA, Qualcomm Ventures and Redpoint Ventures. Connect with MapR on LinkedIn and Twitter.

OpenText
www.opentext.com/what-we-do/products/analytics

OpenText Analytics Suite provides powerful analytics, reporting, and data visualization technology that organizations need to build high-scale, data-driven applications. The OpenText integrated analytics platform enables IT to deliver managed self-service capabilities for business users. Analytics Suite 16 features common, shared services such as single sign-on, single security model, common access, and shared data, and can be deployed on-premises, in the cloud, or in a hybrid environment.

Snowflake Computing
snowflake.net

Snowflake Computing, the cloud data warehousing company, has reinvented the data warehouse for the cloud and today's data. The Snowflake Elastic Data Warehouse is built from the cloud up with a patent-pending new architecture that delivers the power of data warehousing, the flexibility of big data platforms, and the elasticity of the cloud—at a fraction of the cost of traditional solutions. The company is backed by leading investors including Altimeter Capital, Redpoint Ventures, Sutter Hill Ventures, and Wing Ventures. Snowflake is headquartered in Silicon Valley and can be found online at snowflake.net.

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

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