research

BEST PRACTICES REPORT

Q2 2020

Data Management for Advanced Analytics

By Philip Russom

Co-sponsored by:

denodo

tdwi

Transforming Data With Intelligence™

research

BEST PRACTICES REPORT

Q2 2020

Data Management for Advanced Analytics

By Philip Russom

Table of Contents

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.

This report is based on independent research and represents TDWI's findings; reader experience may differ. The information contained in this report was obtained from sources believed to be reliable at the time of publication. Features and specifications can and do change frequently; readers are encouraged to visit vendor websites for updated information. TDWI shall not be liable for any omissions or errors in the information in this report.

About the Author

PHILIP RUSSOM, Ph.D., is senior director of TDWI Research for data management and is a well-known figure in data warehousing, integration, and quality. He has published more than 600 research reports, magazine articles, opinion columns, and speeches over a 23-year period. Before joining TDWI in 2005, Russom was an industry analyst covering data management at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and consultant, was a contributing editor with leading IT magazines, and a product manager at database vendors. His Ph.D. is from Yale. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

About TDWI Research

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management 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 data management 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.

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 business intelligence issues. Please contact TDWI Research Director Philip Russom (prussom@tdwi.org) to suggest a topic that meets these requirements.

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 responded 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, and Rod Gosser.

Sponsors

Datastax, Denodo, Hitachi, Matillion, Oracle, SAP, Snowflake, and TIBCO sponsored this report.

Research Methodology and Demographics

Report Scope. This report makes three assumptions about data management for advanced analytics. First, there are many forms of advanced analytics, including data mining, text mining, natural language processing, statistical analysis, graph, machine learning, and predictive analytics. Second, each form of analytics—and sometimes each individual analytics solution—has requirements for how data must be sourced, collected, integrated, improved, remodeled, stored, and presented. Third, for the greatest business impact, users must demand data that's managed and prepped according to the specific requirements of each analytics use case.

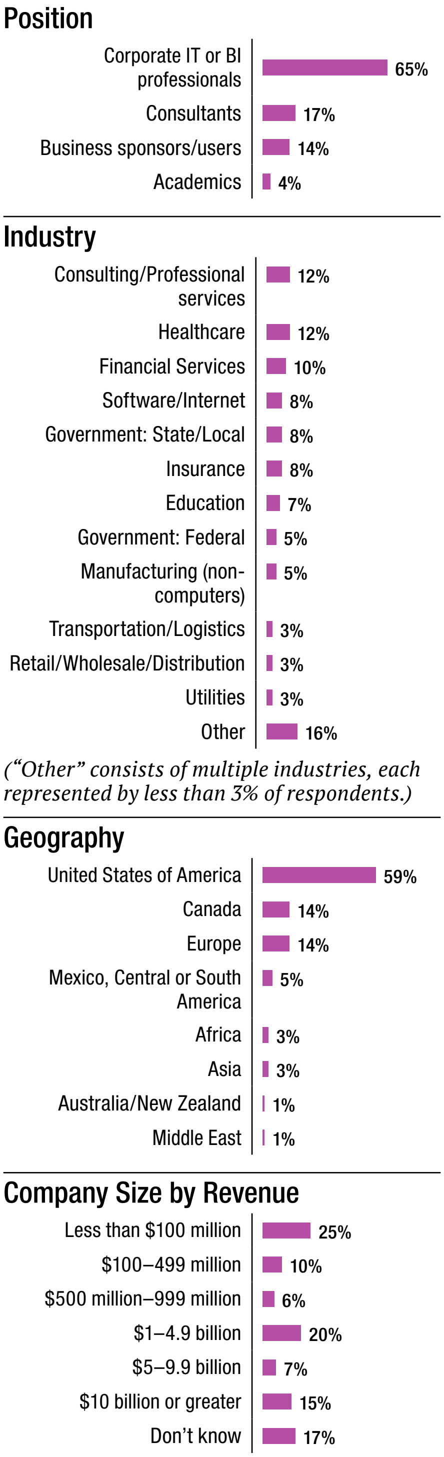
Audience. This report targets business and technical managers who are responsible for creating effective data-driven programs that involve advanced forms of analytics. This report sorts out the data management requirements for common forms and use cases of advanced analytics.

Survey Methodology. In January 2020, TDWI sent an invitation via email to the data management professionals in its database, asking them to complete an online survey. The invitation was also distributed via websites, newsletters, and publications from TDWI and other firms. The survey drew responses from 210 survey respondents. From these, we excluded respondents who identified themselves as vendor employees, and we excluded incomplete responses. The resulting complete responses of 155 respondents form the core data sample for this report.

Research Methods. In addition to the survey, TDWI Research conducted telephone interviews with technical users, business sponsors, and recognized experts. TDWI also received product briefings from vendors that offer products and services related to the best practices under discussion.

Survey Demographics. The majority of survey respondents are IT or BI/DW professionals (65%). Others are consultants (17%), business sponsors or users (14%), and academics (4%). We asked consultants to fill out the survey with a recent client in mind.

The respondent population is dominated by industries in consulting (12%), healthcare (12%), and financial services (10%), followed by software/internet (8%), state/local government (8%), and insurance (8%). Most survey respondents reside in the U.S.A. (59%), Canada (14%), and Europe (14%). Respondents are distributed across all sizes of organizations, though there are fewer very large ones.



Demographics based on 155 respondents.

Executive Summary

"Garbage in" leads to "garbage out," even with modern data management and analytics.

Modern enterprises are expanding their analytics programs to improve their ability to make fact-based decisions, plan for an uncertain future, compete on analytics, and grow customer accounts. These high-value business goals require advanced forms of analytics, which in turn demand use-case-appropriate data integration, data platforms, and other data management (DM). Without the right data in the right format on the right platform, critical and expensive efforts in advanced analytics (AA) have little or no business value.

DM for AA is complex due to the extreme diversity of AA forms and DM options.

Addressing this problem is challenging because there are many forms of AA, including statistical analysis, data mining, clustering, graph, neural net, text mining, natural language processing, artificial intelligence, machine learning, and predictive analytics. Likewise DM includes many types of databases and other data platforms plus tools for integration, quality, metadata, event processing, and so on. To sort this out, this report defines *data management for advanced analytics* (DM for AA), which tailors established and emerging DM best practices and techniques to specific forms of AA, thereby raising the precision, productivity, and business value of analytics.

DM for AA is all about mapping DM options to AA requirements.

The secret to successful DM for AA is to match a combination of DM platforms and tools to each specific use case for AA. For example, for analytics approaches that demand massive data volumes (e.g., mining, clustering, statistics), users tend to deploy Hadoop or a cloud-based DBMS for their analytics data. Some analytics tools run best "in database," which means you must acquire a data platform that supports the form of in-database analytics you need. Real-time analytics requires tools for real-time data ingestion. To succeed with self-service analytics, you need solid business metadata and possibly a data catalog.

DM for AA has compelling benefits and minimal barriers.

Most people responding to this report's survey (94%) find DM for AA to be an opportunity because it increases the usefulness, accuracy, and business value of advanced analytics. The leading benefits of DM for AA include improvements to operations, analytics outcomes, DM upgrades, and real-time data and analytics. The downside is that DM for AA involves more work and expertise for data management professionals plus a longer list of data platforms and tools to acquire and manage. Potential barriers to successful DM for AA may arise in governance, architecture, skills, and DM infrastructure. Given its numerous compelling benefits, most survey respondents consider DM for AA to be extremely important (79%).

Cloud-based data, data platforms, and DM tools are established and growing.

Users perform DM for AA with a wide range of data platforms and tools, both on premises and in the cloud. These include data warehouses (81% on premises, 33% cloud), data integration platforms (68% on premises, 32% cloud), data lakes (43% on premises, 29% cloud), and analytics tools (81% on premises, 42% cloud). These are currently prominent on premises yet well established on cloud platforms. TDWI expects the "cloud gap" to shrink as cloud providers and software vendors raise the maturity of their offerings. Furthermore, survey data suggests that data volumes for AA managed on cloud platforms will *quadruple* within three years. Other tools important for DM for AA include those for data semantics, data virtualization, self-service data, and real-time integration for real-time analytics.

Machine learning and self-service are hot, and so are highlighted in this report.

This report canvasses current and future data management strategies and best practices, then links combinations of these to the leading forms of advanced analytics. The focus is on data management more than analytics. The intention is to help DM and AA professionals and their business counterparts achieve greater success and business impact. Two of the hottest growth areas in AA today are self-service data practices and machine learning, and so this report concludes with detailed discussions of DM requirements for these.

Introduction to Data Management for Advanced Analytics

Defining Advanced Analytics, Data Management, and Their Relationship

Advanced analytics is a collection of multiple user practices and tool types supporting techniques for data mining, text mining, natural language processing (NLP), statistics, clustering, graph, artificial intelligence, machine learning, predictive analytics, self-service, data visualization, and others. In other words, we say "analytics" as if it is a single discipline, whereas in reality it is a collection of many distinct practices. In fact, each approach to analytics has its own focus, abilities, performance characteristics, value proposition, and—as we will discuss in detail in this report—data requirements.

There are many forms of analytics, and each has its own goals, use cases, and technical requirements.

Note that (in this report and its survey) advanced analytics does not include reporting, dashboards, and online analytical processing (OLAP). Similarly, this report distinguishes between reporting and analytics because the two track business entities differently, produce different outcomes, use different enabling technologies, and serve different user constituencies. Even so, this report will mention reporting, dashboards, and OLAP occasionally, because—like analytics—they also have unique data requirements that affect their efficacy.

Data management concerns many diverse product types, technologies, and user practices, all contributing to the successful handling of data. These group into two broad areas:

Both halves of data management must be adapted to the rigors of advanced analytics.

1. **Data integration** captures and repurposes data for applications in reporting, analysis, and operations, using practices and tools for ETL, ELT, data prep, data quality, data virtualization, event processing, metadata management, and data cataloging.
2. **Data platforms** are where data is stored and managed to be provisioned for a wide range of applications. Most data platforms are some kind of database management system (DBMS)—or simply "database." These include older brands of relational DBMSs, newer cloud-based DBMSs, columnar DBMSs, NoSQL databases, and in-memory databases. Non-DBMS data platforms include Hadoop, various file systems, and bare-metal storage. Most of these can be deployed on premises, on cloud platforms, or in a hybrid combination.

As we shall see, each technique for advanced analytics—and sometimes each individual analytics solution—can demand a unique combination of the above-described data integration tools and data platforms.

Data management for advanced analytics is an emerging practice that seeks to raise the targeting, accuracy, and business applicability of analytics outcomes by adjusting generic DM practices to adapt to the unique needs of each analytics technique and solution. It is necessary to coordinate DM and AA in a new and tighter fashion because each approach to AA has its own peculiar combination of data requirements. More to the point, satisfying the requirements leads to a targeted solution that end users will consider a success, whereas leaving requirements unaddressed leads to a solution with limited impact or precision that end users will consider a failure.

DM for AA arose from adapting DM to AA. DM for AA is now a critical success factor for an analytics program.

The Assumptions of This Report

Building on the above definitions, we can now summarize our positions:

- Advanced analytics is not one thing. It covers many approaches, and each has its own purpose, value proposition, use cases, and enabling technologies. Knowing the characteristics of each is fundamental to making good decisions about which to use when.
- Each approach to advanced analytics has a collection of requirements for data management. Fully satisfying requirements leads to a successful solution. Data and analytics professionals who ignore the requirements risk failure.
- Data management for advanced analytics brings the disciplines of AA and DM closer together to ensure that AA solutions get data from the most appropriate sources, containing rich information about business entities of interest, in the best schema for the AA tools being used, with an acceptable level of quality delivered at the right time through an optimal interface.
- To achieve DM for AA and the maximum business value it guarantees, you must tailor your DM best practices and tool usage to the needs of individual analytics solutions. In other words, you cannot perform DM for AA in a single way and expect all implementations of advanced analytics to yield equally useful and accurate outcomes.

Common Pairings of AA Approaches and DM Infrastructure

Each form of AA maps to forms of DM.

To get a better idea of how DM for AA works, consider several scenarios that bring analytics and data management together. For example, the large volumes of human speech captured in text files that are required for NLP differ radically from the lightly standardized tabular data required for self-service data practices. As another example, algorithms for data mining, statistics, and machine learning work well with unstructured or inconsistently structured data of poor quality and no metadata, whereas data warehouse analytics based on time series, hierarchies, or dimensions demand ruthlessly structured, cleansed, and documented data.

Some analytics approaches have multiple sets of data requirements. For example, machine learning involves a complex development and production life cycle; across the life cycle, exploratory data, learning data, training data, and production data are integrated and managed differently. Similarly, in a unified self-service analytics process, the end user moves through data browsing, data prep, visualization, analytics, operationalization, and collaboration; the data requirements of each step vary slightly, yet the process requires that all steps share the same metadata, data interfaces, GUI, and security or governance controls. In addition, mature users regularly deploy two or more approaches to analytics to answer a single business question because each approach provides a different insight; the challenge is to satisfy the data requirements of all coordinated approaches.

Understanding AA/DM pairings is the first step in designing your DM for AA infrastructure.

Finally, a knowledge of the data requirements for analytics can guide the selection of tools and platforms. For analytics approaches that demand massive data volumes (e.g., mining, clustering, statistics), users tend to deploy Hadoop or a cloud-based DBMS for their analytics data. Some analytics tools run best "in database," which means you must acquire a data platform that supports the form of in-database analytics that the tool or use case requires. For real-time analytics, you will need tools for real-time data ingestion. To succeed with self-service analytics, you need solid business metadata and possibly a data catalog.

Real-World Use Cases of Data Management for Advanced Analytics

The vast majority of people responding to this report's survey feel that they know what DM for AA is and does. (See Figure 1.) The high percentage confirms that DM for AA is a real thing. As we'll see later, the majority of survey respondents already have hands-on experience managing data specifically for advanced analytics. This validates that DM for AA is, indeed, well established in the real world.

Most users are familiar with DM for AA and have done it.

Do you believe you know what data management for advanced analytics (DM for AA) is and does?

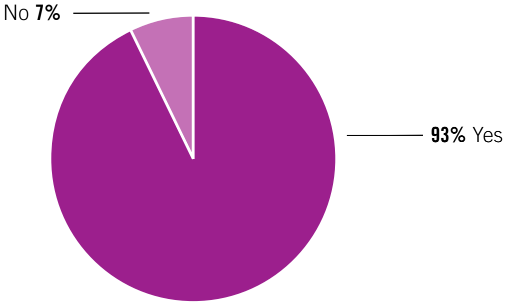


Figure 1. *Based on 155 respondents.*

Business Functions Augmented by Advanced Analytics

To quantify the presence of analytics in practical business functions, this report's survey asked: For what business functions has your organization deployed applications of advanced analytics? (See Figure 2.) Responses show many business functions are supported by some form of analytics.

Users surveyed use analytics to augment business functions in management, operations, marketing, and finance.

Decision making and strategic planning. It should be no surprise that users' primary use of analytics is to enlighten daily business management decision making (55%) and (to a lesser degree) business management strategic planning (38%).

Operations (53%). The processes and decisions of many operational business functions today are guided by analytics, including those for call center and customer service (25%), human resources (23%), and supply chain or procurement (18%).

Marketing (47%). As we'll see later in the discussion of Figure 3, marketing is one of the most analytics-driven business functions today, along with sales (38%), research and development (26%), and product management or brand management (21%).

Finance (41%). Decades ago, the finance department (previously called accounting) was typically the only business function augmented with reporting and analytics. Finance is still an active user, funder, and sponsor of analytics, though eclipsed by operations and marketing in some enterprises.

Other (14%). A few survey respondents selected "Other" as an answer. They cited a number of business functions augmented by advanced analytics, such as compliance, resource management and surveillance, and water and wastewater operations. Only a handful of respondents (3% of total) say their organization has not yet deployed analytics for specific business functions.

For what business functions has your organization deployed applications of advanced analytics? Select all that apply.

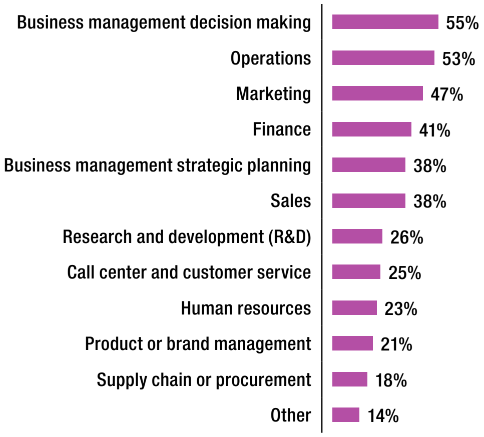


Figure 2. *Based on 617 responses from 155 respondents; 4 responses per respondent on average.*

Analytics Applications in Production

Users surveyed have analytics in production for traditional decisions, customer activities, operations, and risk.

To get a sense of the kinds of analytics applications organizations have in production, this survey asked: What advanced analytics applications has your organization deployed? (See Figure 3.) Responses show that production analytics are common, especially with finance and anything concerning customers. Furthermore, responses indicate that embedded analytics is on the rise.

Traditional analytics. Classic applications for analytics are still highly prevalent, such as those based on budgeting, forecasting, and financial planning (53%).

Customer analytics. This includes many sales, marketing, and customer-oriented tasks and analytics applications, such as customer-base segmentation (41%), customer profitability (35%), customer-churn detection and prediction (34%), marketing campaign design and execution (29%), sentiment analysis (19%), and multichannel marketing (16%).

Operationalized and embedded analytics. A surprisingly high percentage of survey respondents report deploying advanced analytics in the form of automatic decisions in operational applications (34%). When analytics is embedded in operations, it often relies on real-time or streaming analytics (28%).

Fraud and risk analytics. This includes fraud detection and prevention (26%), which usually depends on anomaly detection (26%). TDWI sees its clients in insurance and financials currently upgrading their fraud and risk analytics so they can move beyond detecting a problem after the fact. Instead, they need to proactively predict the problem and stop it before it occurs.

Other (15%). A few survey respondents selected “Other” as an answer. Their typed responses include several applications for advanced analytics, such as risk prediction, dynamic pricing, automatic classification of text assets, income simulation, and customer retention probability. A handful of respondents (4% of total) say their organization does not yet have analytics in production.

What advanced analytics applications has your organization deployed? Select all that apply.

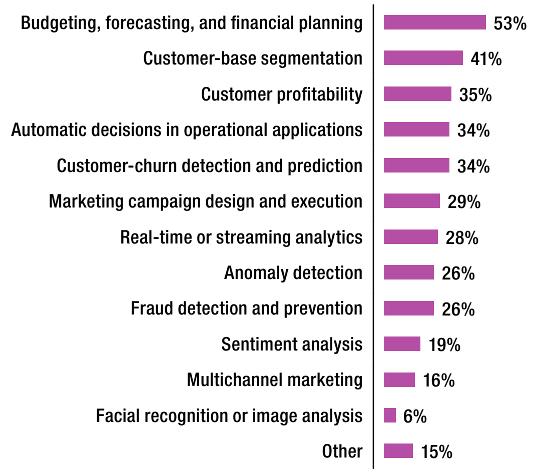


Figure 3. *Based on 560 responses from 155 respondents; 3.6 responses per respondent on average.*

Perceptions of DM for AA and Related Disciplines

The goal of this report and its survey is to convince technology and business users that analytics is a collection of approaches, each approach has distinct data requirements, and satisfying data requirements is a critical success factor for analytics. However, one of the barriers to these realizations is that some data professionals mistakenly believe they can keep managing data in exactly the same way as for traditional reporting and data warehouses and that will suffice for advanced analytics.

Reporting and analytics are different practices with different technical requirements.

TDWI's position is that business reporting and report-oriented data warehouses are not going away because almost no enterprises can function without them. Therefore, you must continue to provision data for them using mature, well-polished best practices. At the same time, you must satisfy the slightly different data requirements of advanced analytics, which may lead you to learn new DM practices such as data prep, pipelining, and orchestration. In other words, you must do it all, both old and new, simultaneously. This puts a strain on your data management team, budget, and infrastructure. However, it is a price you have to pay for high-quality analytics—as well as high-quality reports and warehouses.

To quantify how data professionals and their business counterparts think of these distinctions, our survey asked respondents to select one "truth value" for each row in a table. (See Figure 4.) Their responses show that most users are fairly enlightened today, but there is still room for attitudinal adjustment.

Most reporting is about regularly tracking known entities and processes (71% true, 18% maybe). There are exceptions, but the numbers, metrics, key performance indicators (KPIs), and charts in most reports or dashboards represent business entities that you know very well, and you know exactly how to interpret quantifications of their states and behaviors. For example, this is true of base measures for inventory levels, realized revenue last month, fourth quarter's actual sales numbers in the western region, a time series of daily production yield on the manufacturing floor, and calculated metrics for departmental productivity, customer profitability, and known manifestations of customer churn. Only 11% consider this statement false, which shows that most survey respondents understand the true nature of reporting.

Most analytics is about discovering unknown facts and relationships (62% true, 25% maybe). This is true of discovery-oriented AA such as that enabled by data mining, graph, or clustering. It is also true of the iterative ad hoc queries typical of self-service analytics practices. Note that 25% of respondents scored this statement as maybe. Perhaps they are thinking of how discovery analytics often feeds into production analytics. For example, early in a machine learning project (and sometimes in statistics projects), a data scientist, analyst, or data professional explores a lot of data until an epiphany leads to a prototype of an analytics model. Once the analytics model is in production, the ad hoc discovery mission that started the project is replaced by a consistent and well-understood analytics algorithm or model. Again, most respondents see and understand the discovery orientation of many—but not all—applications of advanced analytics.

Reporting and analytics are different practices (79% true, 9% maybe). As just discussed, reporting's regular tracking of well-known entities and simple quantifications are quite different from analytics's discovery-oriented exploration of data which leads to rich correlations of disparate data points and predictive modeling. Other differences include enabling technologies, tool types, and the types of end users consuming the data-driven products of these practices. Even a single user tends to "wear different hats" when consuming both reports and analytics.

Despite their differences, reporting and advanced analytics have much in common, too. In particular, user organizations regularly use a single data integration and quality toolset for both. Furthermore, reporting and analytics (and many operational applications, too) may employ the same kinds of data platforms, typically relational DBMSs on premises and/or in the cloud, or an alternative such as Hadoop. In these cases where DM infrastructure is shared, the data integration tools and data platforms are used differently for these different use cases. That's one of the points of DM for AA: you cannot manage data only one way and use DM infrastructure only one way and credibly expect to satisfy all requirements for all reporting and analytics use cases. To ensure success, you must tailor DM practices and infrastructure to individual use cases.

Advanced analytics involves several user practices and tool types (80% true, 15% maybe). It is convenient to have the categorical label "analytics," which we say as if it were a single thing. In practice, advanced analytics is a long list of many techniques, each with its own tool types, user best practices, and data requirements.

Data management requirements vary across diverse forms of advanced analytics (75% true, 17% maybe). Given the diversity of analytics approaches plus the unique abilities, performance characteristics, and intended outcomes of each, it is inevitable that data requirements are also highly diverse for analytics approaches as a whole.

Please select one “truth value” for each row of the following table.

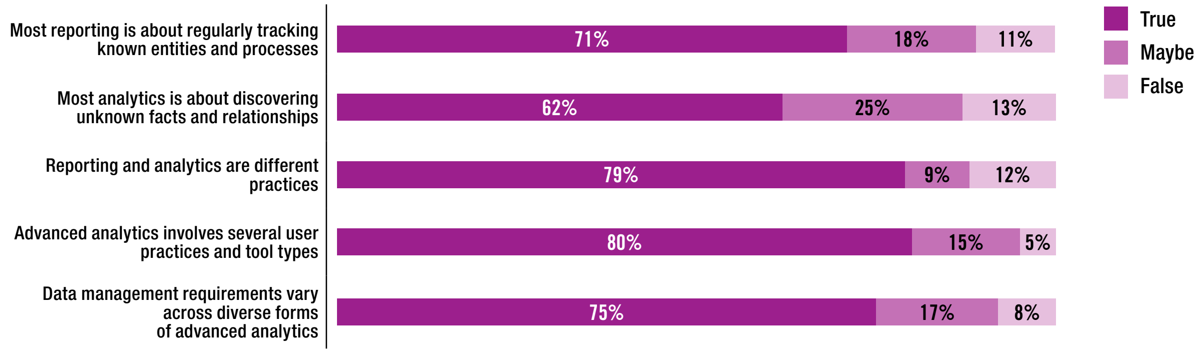


Figure 4. Based on 155 respondents.

Traditional data warehousing is a great fit for reporting but not so much for AA.

To further explore respondents’ perceptions of data management and advanced analytics, our survey asked respondents to compare data management for a traditional report-oriented data warehouse to data management for advanced analytics. (See Figure 5.)

Very few respondents see the situation as purely black or white. At one extreme, almost no one considers DM for report-oriented warehouses and advanced analytics to use identical practices and tools (1%). At the other extreme, few consider them totally different (9%).

The vast majority see an overlap of the two practices. Some see a large overlap, where the two are mostly the same with some differences (37%). However, a slight majority see a small overlap, where DM for report-oriented warehouses and analytics are mostly different with some similarities (53%).

Compare data management for a traditional report-oriented data warehouse to data management for advanced analytics.

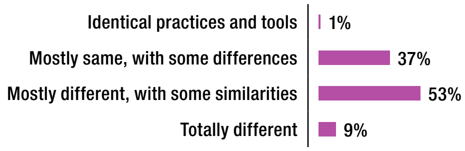


Figure 5. Based on 155 respondents.

The issue raised by the survey question in Figure 5 is that many owners of legacy data warehouses are under pressure to also support a broadening range of analytics data requirements. This begs important questions:

- **Should we stretch the existing warehouse to support DM requirements for AA?** TDWI has seen organizations accomplish this successfully, but success depends on having a powerful DMBS platform already in place under the legacy DW that can scale easily and perform well with diverse concurrent workloads. Beware that this strategy is extremely expensive when implemented with a large configuration of an on-premises MPP DBMS.
- **Should we build a new database optimized for AA?** Most legacy data warehouses are in excellent condition, with quality data and solid platforms. Instead of upsetting this apple cart, the trend is to leave the legacy warehouse intact under the assumption that “it ain’t

broke, so don't fix it." In this strategy, users stand up one or more new data platforms (e.g., Hadoop, data lake, cloud database, columnar database) then optimize them for AA and ensure integration with the legacy warehouse.[1] This drives up the complexity of the warehouse architecture, but DW professionals have a track record of success with such complexity.

Note that TDWI is not making a recommendation here, nor are we saying that DWs cannot handle AA. (In fact, TDWI's position is that a data warehouse can support AA or anything you want as long as you modernize it accordingly and you are willing to pay the bill.) We are just pointing out the issues so users can make informed decisions—based on their unique situations and directions—about their strategies for modernizing a legacy data warehouse in situations where AA support is a high priority.

> **USER STORY REPORTING AND ANALYTICS ARE DIFFERENT.**
>
> "We don't have a formal definition of analytics where I work," said a report designer at a large media firm, "but I think of reporting as spitting out data result sets for refreshing reports. Each report was designed for a specific use or user and carefully vetted so that everyone understands what the numbers represent.
>
> "Analytics is different because it's about creating an understanding of entities represented in the data. You have to think beyond the numbers and ask, 'What does it mean to our organization?' You can do this with tools, but for us today analytics is usually mental, verbal, and collaborative, and we go through a lot of data set versions, repeatedly pulling data from many systems, until we reach consensus on that understanding.
>
> "We know we're held back by this manual and somewhat haphazard process. We've founded an analytics program so we can do it right in the future."

Benefits and Barriers for DM for AA

DM for AA: Problem or Opportunity?

Almost all survey respondents see the upside of DM for AA.

To test whether DM for AA is worth the effort, this report's survey asked: Which of the following statements best represents your view of data management for advanced analytics (DM for AA)? (See Figure 6.)

The vast majority of respondents (94%) consider DM for AA an opportunity. Firms and other organizations are deepening their investments in analytics as they expand their analytics programs. Tailoring data management to the needs of new forms of analytics is an opportunity to increase the value of these investments.

A tiny minority (6%) consider DM for AA a problem. Growing data management infrastructure and increasing team workloads to address the needs of new forms of analytics will incur time, costs, and other investments. However, the investment in DM practices and infrastructure has a return in the form of positive impact on the business.[2]

[1] For details about this practice, read the 2018 *TDWI Best Practices Report: Multiplatform Data Architectures*, online at tdwi.org/bpreports.

[2] For a discussion of how data strategies should support multiple forms of analytics, see the 2020 *TDWI Checklist Report: Data Strategies for Accelerating the ROI of Analytics*. Available online at tdwi.org/checklists.

Which of the following statements best represents your view of data management for advanced analytics (DM for AA)?

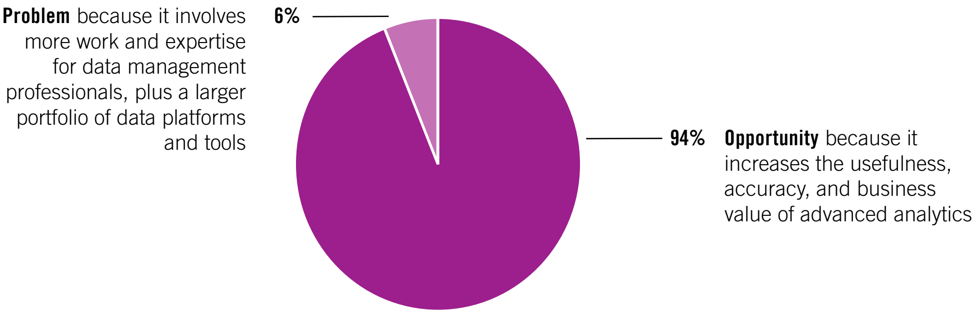


Figure 6. *Based on 155 respondents.*

Benefits of DM for AA

In the perceptions of survey respondents, DM for AA offers several potential benefits. (See Figure 7.) A few areas stand out in their responses:

Better operations via analytics support. Top ranked by survey respondents, DM for AA has the potential to yield better operational decisions due to better data access (68%). The focus on operational improvements via analytics is echoed in other responses, such as the embedding of AA algorithms in operational applications to automate decisions (31%), predictive models used more often and broadly (45%), and analytics for predictive maintenance (15%).

Better analytics outcomes. The second most likely benefit of DM for AA (based on survey responses) is better strategic analytics and planning (65%). The consensus is that DM for AA improves existing analytics applications (20%) and enables a wider range of applications of advanced analytics (30%), smarter recommendations for users and customers (27%), and more accurately targeted customer engagement and personalization (24%).

DM for AA upgrades data infrastructure. Many analytics programs are under pressure to broaden data democratization and enable more users to apply data (51%), typically in support of self-service analytics practices. Upgrades of DM for AA infrastructure result in more actionable and more focused information (48%), where more data assets will be fully leveraged via analytics (37%). In general, DM for AA modernizes a mature data management infrastructure (26%) such that all data-driven applications benefit when data management requirements are taken more seriously.

Real-time analytics based on real-time data. Another pressure point that organizations are feeling is the need to bring the timing of business reactions closer to real time. When DM for AA upgrades real-time data management, the results include more timely decisions (45%), real-time process optimization enabled by analytics (28%), and situation awareness and alerting that is analytics-driven (24%).

The leading benefits of DM for AA include improvements to operations, analytics outcomes, DM upgrades, and real-time data and analytics.

If your organization were to tailor data management more specifically for advanced analytics, what would its leading BENEFITS be? Select seven or fewer.

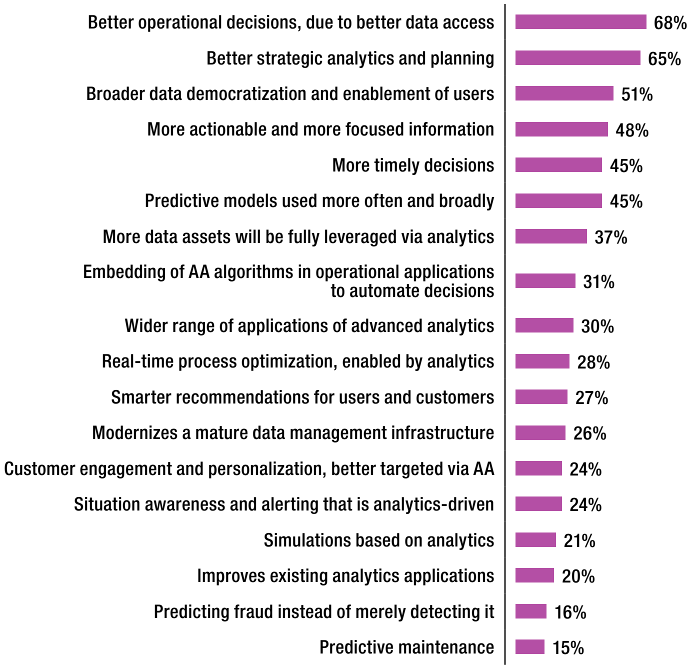


Figure 7. *Based on 965 responses from 155 respondents; 6.2 responses per respondent on average.*

Barriers to DM for AA

Barriers to successful DM for AA may arise in governance, architecture, skills, and DM infrastructure.

In the perceptions of survey respondents, DM for AA presents a few potential barriers. (See Figure 8.) A few areas stand out in their responses:

Issues in governance, privacy, and compliance. The leading barrier to successful DM for AA is data governance. Even organizations experienced with governance are hard-pressed to modernize it to cover the additional data platforms and analytics use cases (60%) that are inevitable with analytics programs. Even so, TDWI sees organizations putting in the time, money, and effort to successfully address broader governance issues as well as data privacy and usage compliance issues (42%).

Complexity of hybrid data architectures. Satisfying the data requirements of multiple forms of analytics—plus reporting, dashboards, OLAP, and other older practices—leads many DM teams to deploy multiple data platforms, each optimized for a particular analytics use case or structure of analytics data. The data platforms may manage data on premises, in the cloud, or both, in systems architectures and data architectures that are distributed and hybrid. The extreme complexity of these modern architectures increases the difficulty of architecting data environments that can repurpose data repeatedly for many AA use cases (44%), modernizing a data warehouse (built for reporting) to handle AA data (48%), and maintaining a single version of the truth that is current and accurate (53%).

The complexity of these architectures is a necessary evil due to their ability to provide the best home for any data and any analytics use case. Besides, TDWI sees organizations embracing hybrid architectures successfully. After all, data warehouse teams and others have worked with distributed data in multiplatform data environments for decades. Today, DM and warehouse professionals apply those skills to a whole new level of complexity, eager to reap the benefits of new data platforms, open source, and clouds.

Inadequate DM for AA skills and headcount. Organizations that are new to advanced analytics are initially held back by a lack of personnel skilled with AA (45%) and/or a lack of DM skills relative to requirements for AA (46%). Even established analytics programs struggle to keep pace with skills and headcount issues. TDWI sees many teams successfully solving these problems with a combination of retraining existing employees, hiring more employees, and engaging consultants who have DM and AA experience.

Miscellaneous small data management challenges. A number of potential barriers to DM for AA may arise in areas of data management, but the low percentages among survey responses mean that relatively few users are concerned. For example, there is some concern over the poor quality of data from traditional sources (34%) and new sources (22%). Keep in mind that DM for AA will entail upgrades and additional tooling—and perhaps replatforming—when integration infrastructure has limited functionality or performance (24%). Handling too many disparate data sources (28%) is a challenge where new sources come online daily, often due to the Internet of Things (IoT), multichannel marketing, or growing partner ecosystems in business-to-business scenarios. Likewise, learning the interfaces and schema of external data sources (19%) is a bit challenging. Note that users are not held back by nontraditional data (17%) or streaming data (5%).

If your organization were to tailor data management specifically for advanced analytics, what would its leading BARRIERS be? Select seven or fewer.

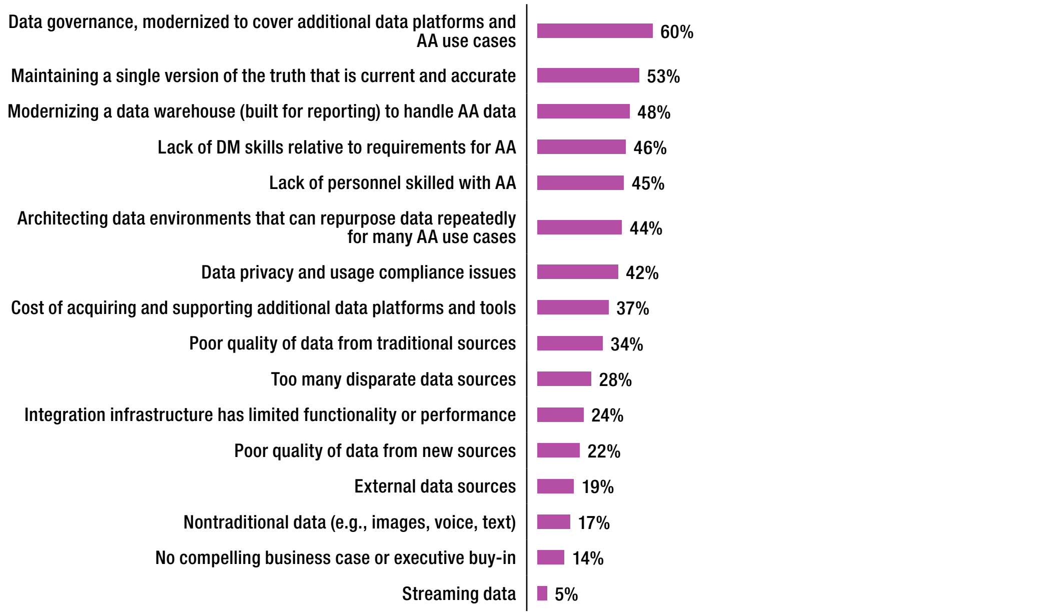


Figure 8. *Based on 836 responses from 155 respondents; 5.4 responses per respondent on average.*

USER STORY DATA LAKES AND MACHINE LEARNING FORM AN EMERGING ANALYTICS ARCHITECTURE.

"In our firm, individual business units have their own IT, which makes it difficult to monitor the firm in a holistic way," said Nadine Schramm, data governance and management technology leader at a large energy company. "To address this, about a year ago we founded our enterprise data platform team. The team's job is to create a view via data that spans across all business units. To enable this view, we are designing an enterprise data model, and its core is already done and in use. We have deployed the core model on a cloud data platform so we can populate it with data drawn from across the enterprise. This is essentially a large data lake, built and optimized for advanced analytics.

"With the data lake as a foundation, we are building a wide range of advanced analytics. For example, our CEO is a strong sponsor of the enterprise data team. He wants machine learning for modeling and predicting power generation fluctuations and outages since that's core to what we do as an energy utility. He also wants analytics that support the continuous improvement of work and resource excellence for both employees and contractors.

"Our plans for machine learning are aggressive, with projects slated for process improvement analytics, value-based predictive maintenance, and various planning and forecasting applications. As an energy company, weather has a massive influence on our day-to-day operations and overall success, so we're aggressively amassing data about weather, maintenance, repairs, outages, and operations to better predict weather events and their impact on our business."

The State of DM for AA

Is DM for AA Important?

To gauge the urgency of DM for AA, this report's survey asked: How important is data management to the success of your organization's programs and applications for advanced analytics? (See Figure 9.)

Almost all organizations consider DM a critical success factor for analytics.

Remarkably few respondents (2%) say that DM is not a pressing issue for analytics. Conversely, we can conclude that data management contributes significantly to use cases in advanced analytics.

The vast majority of respondents (98%) recognize the importance of DM. Many feel that data management is extremely important (79%) to the success of analytics, while others see it as moderately important (19%).

How important is data management to the success of your organization's programs and applications for advanced analytics?

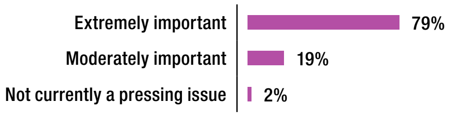


***Figure 9.** Based on 155 respondents.*

Why Is DM for AA Important?

To answer this, the survey asked an open-ended question: In your own words, why is improving data management specifically for advanced analytics important (or not important)? Respondents' typed comments reveal a number of use cases, needs, and trends as seen in the representative excerpts reproduced in Figure 10. Note that the people quoted work in many industries and geographic regions. Clearly, the relationship between data management and advanced analytics is top of mind for data professionals and their business sponsors in many contexts worldwide.

Users have many good reasons for considering DM important to successful analytics.

In your own words, why is improving data management specifically for advanced analytics important (or not important)?

- "Good DM means good data quality that is 100% related to AA quality and success." – consulting, principal, Europe
- "Data management increases the value of our data assets." – financial services, IT director, U.S.
- "It adds tremendous value to the business. We are striving towards a data-driven culture. AA is appropriate to achieving that goal." – federal government, director of analytics, U.S.
- "DM needs to provide quality foundations to reduce the risk of AA returning incomplete or inaccurate results." – insurance, data architect, U.S.
- "To avoid manual processes or cumbersome integrations that quickly become obsolete." – federal government, consultant, Canada
- "To gain efficiency, quality, agility, and speed; to deploy multiple analytical applications effectively." – financial services, chief data officer, Argentina
- "Without adequate DM and quality of the data, AA will not be trustworthy and used in production." – consulting, chief executive officer, Europe
- "To maximize the investment that is made in advanced analytics." – hospitality, consultant, U.S.
- "In order to have larger samples of data both structured and unstructured." – education, BI manager, Europe
- "Data management allows for more streamlined data access by end users as well as real-time advanced analytics." – insurance, BI and analytics manager, Canada
- "Sound data management throughout the life cycle is essential to ensuring accurate analytics insights and promoting better decisions." – healthcare, consultant, U.S.
- "Garbage in = garbage out." – retail/wholesale, BI manager, U.S.

Figure 10. *Drawn from the typed responses of 132 respondents.*

Most Survey Respondents Have Experience with DM for AA

To sort survey respondents according to their exposure to DM for AA, our survey asked: Do you personally have direct experience managing data specifically for advanced analytics? (See Figure 11.)

Two-thirds of organizations surveyed are already managing data for some form of analytics.

The majority of survey respondents have direct experience with DM for AA (64%). This high percentage reveals how pervasive advanced analytics has become. However, why the massive response?

Recent research from TDWI shows that only 35% to 40% of organizations surveyed are following through with their plans for analytics programs.[3] The low percentage is due to the difficulties of acquiring budget, personnel, and skills for new analytics programs. Even established programs have difficulty growing due to the scarcity of experienced employees and consultants for advanced forms of analytics. Despite the challenges, half of organizations recently surveyed have multiple analytics models and other solutions in production. Getting analytics off the ground is problematic, but it can be done in a sustainable way.

Do you personally have direct experience managing data specifically for advanced analytics?

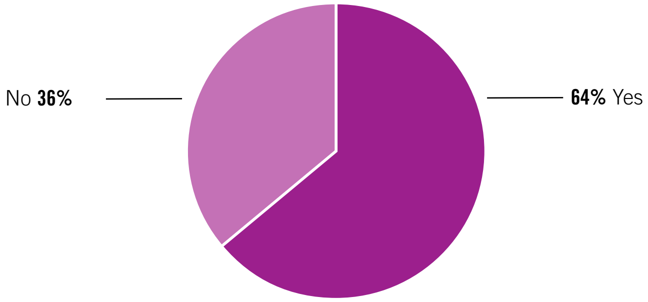


Figure 11. *Based on 155 respondents.*

Note that this report's survey branched according to how respondents answered the question in Figure 11. Respondents answering "yes" (64%) were presented with detailed questions about technologies, teams, and best practices involved with data management for advanced analytics, as seen in many of the following figures.

Data management and advanced analytics are like other IT disciplines. There are successful programs and a few programs that fail outright. However, success and failure are more often a matter of degree in that some aspects of a program succeed while other aspects fail. The program continues and users improve the lackluster parts as they go.

To get a sense of which aspects of DM for AA are currently succeeding or failing, our survey presented two open-ended questions that allowed respondents to describe their successes and failures in their own words (see Figures 12 and 13 respectively). Note that these questions were posed to respondents who have direct exposure to DM for AA, so they speak from real-world experience.

DM for AA Successes

Efforts in DM for AA are seeing impact in DM, AA, marketing, and operations supported by analytics.

Figure 12 assembles several representative comments about aspects of DM for AA that are succeeding today. Most comments collected by the survey directly address DM, AA, marketing, or operations; Figure 12 organizes the comments into these categories. The survey population is dominated by people who work in decision-making disciplines, so it's no surprise that their successes primarily concern reporting, analytics, data warehousing, and data integration, with additional successes for many types of operational applications and their analytics needs. Clearly, respondents' successes prove that DM for AA is real, it works, and it provides business value.

[3] For more information, see the 2018 *TDWI Best Practices Report: Practical Predictive Analytics*, online at tdwi.org/bpreports

Briefly list some areas where data management tailored specifically to advanced analytics has SUCCEEDED in your organization.

Data management

- "DM provides a data lake for AA functions"
- "Established a data lake with success querying various types of data"
- "Timelier and more accurate streaming data and data lakes"
- "DM supports a cloud data pipeline for AA compute-intensive functions"
- "Customer data management and operations data management"
- "Consolidation of legacy databases into fewer systems of truth"
- "Deployed a warehouse/lake hybrid for analytics"

Advanced analytics

- "Machine learning and artificial intelligence program started"
- "Success with specific applications rather than general capabilities"
- "Predictive models for staffing"
- "Analytics for medication comparative effectiveness"
- "Risk predictive models based on claims history"
- "Analytic credit scoring beyond financial data"
- "Fraud and anomaly detection"

Marketing analytics

- "Predictive modeling for customer churn, engagement, and classification"
- "Machine learning-driven models for forecasting demand"
- "Customer behavior and sentiment analysis"
- "More accurate party identification and profiling"
- "Marketing campaign designs based on analytics outcomes"

Operations supported by analytics

- "Predictive models for financial forecasting; audit and tax analytics"
- "Automation efficiency with image analysis"
- "Prescriptive analytics in customer domain; predictive analytics in MRO domain"
- "Predictive maintenance, not just preventive maintenance"
- "University operations—predicting enrollment and show rates"
- "Real-time census to better manage bed capacity at hospital"

Figure 12. *Drawn from the text responses of 99 respondents who have DM for AA experience.*

DM for AA Failures

Failures occur within single areas of DM for AA and so can be understood and remedied easily.

Figure 13 assembles representative comments about aspects of data management and advanced analytics that have failed in some organizations. The point of the list is to inform you of potential potholes you should steer around en route to DM for AA success.

Note that none of these is a total failure. Instead, they are partial failures, each isolated to an aspect of DM for AA such that the failures can be corrected and expanded for long-term success. Also note that few types of failures repeat in respondents' comments, meaning that there are no recurring points of failure in the DM for AA paradigm. Furthermore, a quarter of respondents (24%) reported no failures, meaning that failure is not universal with DM for AA.

Briefly list some areas where data management tailored specifically to advanced analytics has FAILED in your organization.

- "Cannot make real-time integration with operational systems work"
- "No value from advanced tools due to low skills and best practices"
- "We couldn't democratize the data warehouse"
- "No support for analytics from business subject matter experts"
- "Architecture and data source management are harder than they seemed"
- "DM team built a Hadoop cluster that was not usable by AA team"
- "Inadequate IT infrastructure"
- "Tying data insights back to specific business needs is unreliable"
- "Poorly defined use cases"
- "Single version of the truth is unlikely with complex architectures"
- "How to cleanse data differently for different analytics models"
- "Building a data lake took too long and was too costly"
- "Understaffed IT cannot update architecture and build pipelines"
- "Integrating analytics data across entrenched business units"
- "Poor performance and big effort to get report data from a data lake"

Figure 13. *Drawn from the text responses of 99 respondents who have DM for AA experience.*

EXPERT COMMENT **LEADERSHIP, PLANNING, AND COLLABORATION CAN MAKE THE DIFFERENCE BETWEEN SUCCESS AND FAILURE.**

Fern Halper is a senior research director at TDWI and a recognized expert in advanced analytics. According to Halper:

Organizations that succeed in evolving their analytics strategy often have similar characteristics that include being goal-driven, empowering, collaborative, transparent, and skills-based. There are a number of strategies that TDWI survey respondents cite as helpful in advanced analytics efforts:

Find the right champion. This doesn't have to be a chief analytics officer (although that can help). We see many successful programs led by VPs and directors of analytics or those aligned with strategic lines of business. If machine learning is aligned with the business strategy, it can grow successfully.

Identify the right project. Start any machine learning effort with a real business problem and clear objectives. Ideally, the objectives are measurable so the organization can quantify the impact of the machine learning project. This will help to articulate the impact of a successful project and help to get others on board to help machine learning grow.

Organize to execute. We often see successful organizations utilizing a center of excellence (CoE) model. They typically start with a few data scientists who can build models. There may also be business analysts involved. As organizations scale the number of models they build and put into production, they often need to put staff in place (such as DataOps, ModelOps, or DevOps) to deal with deployment and model monitoring.

Make sure to collaborate. TDWI surveys often cite collaboration as a top best practice. For instance, on the technical front in machine learning efforts, IT may be responsible for data engineering and DevOps, while the business may be responsible for building a machine learning model or owning the effort. The more the organization can work together from the get-go, the better the outcome.[4]

DM for AA Tool and Platform Requirements

Tools, Techniques, and Platforms Used in DM for AA Today

The list of tools and platforms involved in DM for AA continues to evolve as users adjust their software investments, usually to include more cloud-based systems. To quantify the mix that users are working with, our survey asked: With DM for AA in your organization, what data and analytics tools, techniques, and platforms are supported today? (See Figure 14.)[5] The overall message from survey responses is that clouds of some kind are being used by almost all user organizations. Furthermore, data is being created, managed, integrated, and used for advanced analytics across hybrid data architectures, as seen in the following examples.

The relational DBMS is still king, but it coexists with many complementary platforms.

Database management systems (DBMSs). Whether on premises, in the cloud, or in a hybrid architecture across both, the relational database management system (91%) retains its hegemony among DBMSs. This is due to the SQL and relational requirements of set-based and self-service analytics. Even so, some users also rely on nonrelational database management systems (42%) and NoSQL database management systems (36%).

Although it is not a DBMS per se, Hadoop (37%) is often used as if it were, especially for algorithm-based analytics programmed in Java, R, and Python. Similarly, TDWI sees increasing numbers of organizations using cloud-based raw storage. Among the approaches to cloud storage (object, file/folder, and block), the one most similar to the DBMS paradigm is object storage on cloud platforms (33%).

[4] Read more of what Fern Halper has to say about successful analytics in the 2020 *TDWI Checklist: Six Success Factors for Getting Started with Machine Learning Across Your Enterprise*, online at tdwi.org/checklists. The Expert Comment above is drawn from the first section of that report.

[5] In the actual survey, the data points shown in Figure 14 were multiple-choice answers for two near-identical questions. For a unified view of the issues, they are presented here as one result set. Two questions in the survey were necessary to make the long list easier for respondents to handle.

Data platforms and tools for DM for AA are mostly on premises today, but also established in the cloud.

Decision-making technologies. The prominent layers of the decision-making technology stack are represented amply in survey results charted in Figure 14, both on premises and on cloud systems. These decision-making technologies include data warehouses (81% on premises, 33% cloud), data integration platforms (68% on premises, 32% cloud), data lakes (43% on premises, 29% cloud), and analytics tools (81% on premises, 42% cloud). Note that each category of decision-making technology is currently more prominent on premises than in the cloud—but not by much. TDWI expects the "cloud gap" to shrink as more users gain confidence in cloud technologies and as cloud providers and software vendors increase the maturity of their offerings.

Data virtualization (37%) usage continues to grow in AA, especially in hybrid and other distributed data ecosystems where data consolidation or movement is not practical, prompting enterprises to build a virtual layer that makes data appear consolidated.

Analytics tools and techniques. Survey responses show that all of the advanced forms of analytics are in use today in respectable amounts. Statistical techniques (78%) were established long before other forms of AA, and statistics continues to be the king of analytics today. In recent years, predictive analytics (65%) has risen aggressively as organizations seek to predict business events and react proactively instead of only detecting and reacting to events after the fact.

Some forms of analytics are tightly related. For example, predictive analytics is often supported by techniques in machine learning (53%) and/or artificial intelligence (32%). These techniques have recently been automated well by tools, such as AutoML tools (20%).

Modern businesses are also seeking to be informed and guided in as close to real time as possible by adopting stream processing and/or analytics (39%). Others seek greater business use and value from unstructured data, as when they use text analytics (37%) or natural language processing (NLP, 31%) in support of sentiment analysis.

Open source (except LINUX, 59%). The operating system LINUX years ago proved the value, performance, reliability, and low cost of open source software. In recent years, various open source tools and platforms have proved themselves useful in data management, including containers (e.g., Docker, 23%), Kubernetes (17%), and Cassandra (10%). Hadoop (37%) is now common in AA and data warehousing, often integrated with Spark (39%) for low-latency microbatching and Kafka (23%) for real-time data.

Data semantics. Almost every task in data management requires metadata management (49%) in one form or another.

Cloud types. As we saw above, every layer of the decision-making technology stack is well established on the cloud. Furthermore, one of the most common project types among TDWI Members today concerns data migration to the cloud. This is because the cloud offers many benefits to DM, namely elasticity for high performance and linear scale, minimal administration, and low cost compared to equivalent deployments on premises. So far, most cloud deployments for AA and other decision-making disciplines have involved a single provider (38%), but many will surely evolve to use multiple providers (25%).

In a related trend, TDWI sees users increasingly looking for third-party cloud providers that have a managed service (29%) for cloud versions of platforms they need, typically a favored brand of DBMS, distribution of Hadoop, analytics platform, or tool for data integration.

Real time. Real-time analytics won't happen without real-time DM tools for streams, IoT data sources, and monitoring (30%) or event processing (32%). In Hadoop and similar open source environments, Kafka (23%) is well established for real-time data ingestion.

With DM for AA in your organization, what data and analytics tools, techniques, and platforms are supported today?

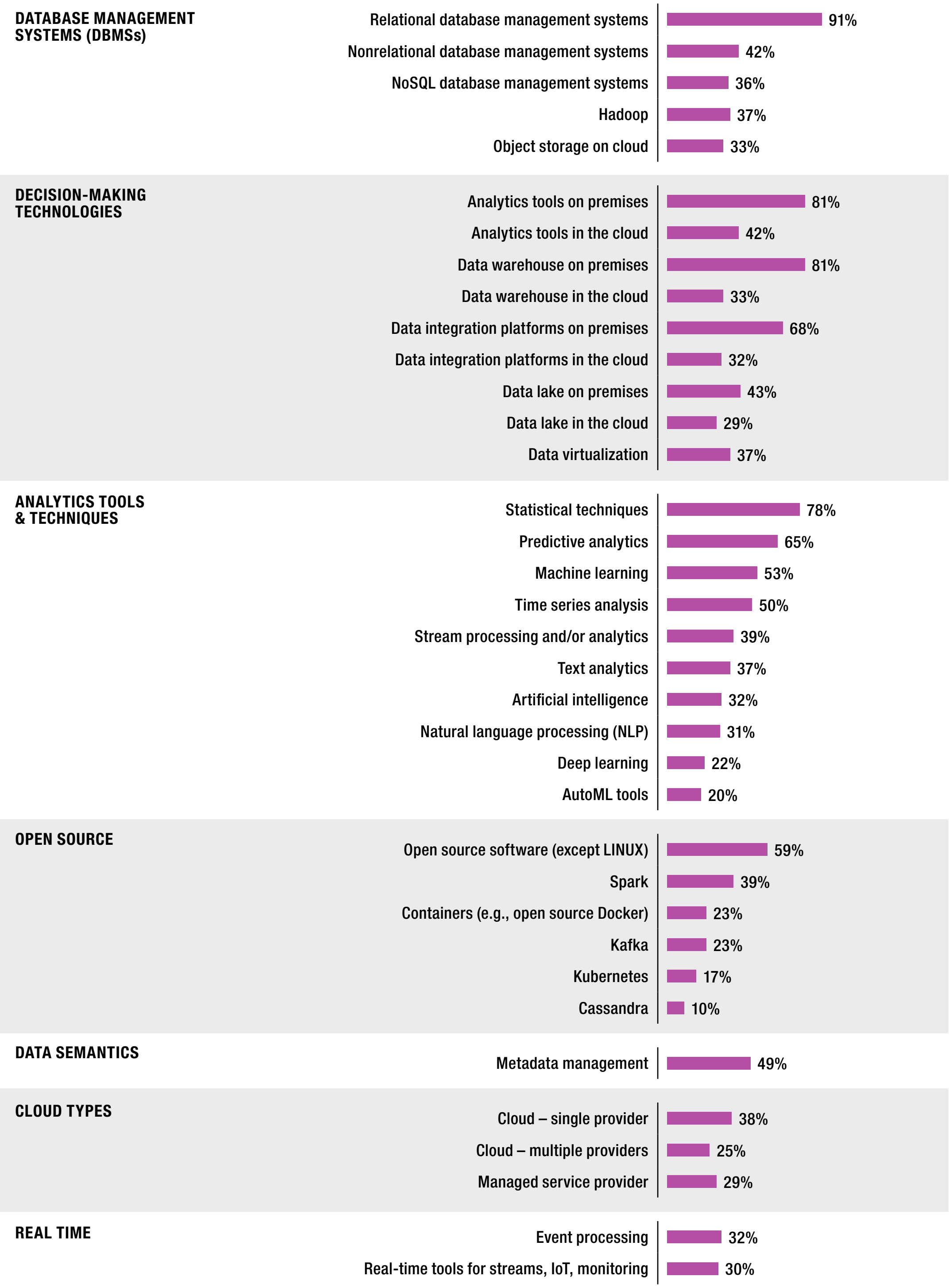


Figure 14. *Based on 98 respondents who have DM for AA experience.*

Data Management Capabilities Critical to AA Success

This report's survey asked: What data management capabilities do you need for successful advanced analytics? The capabilities chosen most often by respondents reveal what experienced users rely on most often today, which in turn suggest priorities for all organizations seeking success with DM for AA. (See Figure 15.)

Users have many options to draw from when assembling DM infrastructure for AA.

Data warehouse (85%). The data warehouse and data integration tied as the highest priorities for DM for AA. This is not a surprise because the two continue to be the core of data management for both reporting and analytics, albeit with modernizations to keep pace with technical advances. Furthermore, the two work hand in hand, interoperating constantly in development and production such that it can be difficult to tell where integration ends and the warehouse begins.

The high priority that respondents give the DW corroborates that it is still highly relevant to AA despite the fact that most DWs were originally designed primarily for reporting. To address this, users regularly modernize their DWs to make them more conducive to AA by complementing and extending the DW with a data lake (52%). A unified warehouse/lake architecture can be deployed on premises, in the cloud, or distributed across both.

In a unified warehouse/lake architecture, the data lake assumes responsibilities for data ingestion, persisting detailed source data for exploratory analytics, and staging data for the data warehouse and other downstream systems. In this scenario, the data lake is an extension of data integration, and it provisions massive volumes of detailed source data for AA, whereas the DW provisions carefully transformed, cleansed, and documented data for reports, dashboards, OLAP, and any analytics that require historical data (e.g., budgeting, forecasting, and strategic planning).

Data integration (85%). This includes traditional approaches to data integration (e.g., ETL, ELT, replication, synchronization, and federation) as well as new approaches such as data virtualization (33%), data pipelining (47%), and orchestration and workflow management (45%). This category also includes data disciplines that are closely related to data integration, with a high priority for AA success placed on data quality (80%).

Data semantics. This is a broad term that encompasses all forms of metadata management (53%), as well as other semantic methods for describing and categorizing data. Technical metadata continues to be the preferred semantic for interfaces among systems, including those that collect data for AA. However, business metadata is increasingly important because it is a requirement of self-service analytics. Furthermore, business metadata may be organized as a business glossary (54%) and may be replaced eventually by modern approaches to the data catalog (53%).

Note that a number of semantic approaches are built on top of technical metadata, the highest priority one being business metadata. Other valuable data semantics built from metadata include impact analysis (37%), data lineage (36%), and data virtualization (33%).

Self-service data practices. As just noted, business-friendly data semantics (business metadata, glossaries, and catalogs) are a high priority for successful self-service analytics. Likewise, other priorities for self-service analytics include easy-to-use self-service data tools (58%) and data prep for simplified integration (58%).

Interface and API management (60%). For the broadest analytics correlations possible, DM for AA should integrate data from many sources at multiple latencies. Also, the number and type of sources and targets are growing in most enterprises, often via new interfaces. New sources and targets relevant to AA include software-as-a-service (SaaS) applications, the Internet of Things (IoT), increasing numbers of marketing channels, growing B2B partner ecosystems, and third-party data providers. This is why interface and API management is a growing discipline in data integration in general, as well as in DM for AA.

People-driven data practices. Data managed for AA has limited business value, governance, and accessibility unless DM for AA supports data sharing functions (49%), stewardship and curation features (44%), and tool features for multiple user types (36%).

Real-time analytics. Enabling this form of advanced analytics requires real-time data interfaces (41%), which may be enabled by in-memory functions (40%), event processing (37%), and edge analytics (13%).

Data services. Though still new (and therefore still relatively low priority for most user organizations), data-as-a-service (DaaS, 25%) and microservices for data (19%) are established and upcoming.

What data management capabilities do you need for successful advanced analytics? Select all that apply.

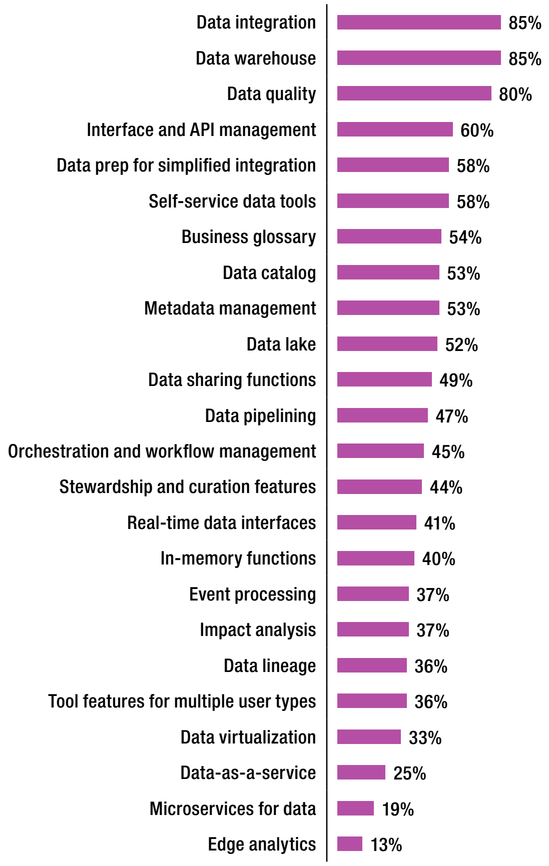


Figure 15. *Based on 99 respondents who have experience with DM for AA.*

Physical Locations of Analytics Data and Its Sources

The recent proliferation of new data platforms and SaaS applications has increased the presence of cloud-based systems in modern enterprises, such that an increasing amount of data “lives” on the cloud systems regardless of where it may have originated or may be going. Data management professionals need to track this trend from a capacity-planning viewpoint—as well as to ensure that the data is where it needs to be for specific use cases in advanced analytics. To get a sense of how the physical distribution of data for analytics is shifting, our survey asked: Which of the following best describes the location of data specifically for advanced analytics, relative to on-premises versus cloud systems, as it is today, as well as in three years? (See Figure 16.)

The primary home of analytics data will shift from on premises to the cloud.

Today, most enterprises surveyed have most of their data on premises. Respondents report having analytics data almost exclusively on premises (49%) or mostly on premises (28%). Conversely, single-digit percentages of respondents report having analytics data on cloud systems mostly (5%) or exclusively (6%) today.

Managers of analytics programs must plan for significantly heavier loads of analytics data in the cloud, but less on premises.

In three years, on-premises systems will no longer be the primary home for analytics data. For example, very few respondents expect their analytics data to be almost exclusively on premises (8%) in three years, and relatively few anticipate it will be mostly on premises (28%). Instead, most respondents think their analytics data will be near equally on premises and cloud (21%), excessively hybrid and mostly cloud (20%), or almost exclusively cloud or multicloud (23%). From the survey data seen in Figure 16, we can estimate that the total volume of data managed for advanced analytics on cloud systems will roughly *quadruple* within three years!

Which of the following best describes the location of data specifically for advanced analytics, relative to on-premises versus cloud systems, as it is today, as well as in three years?

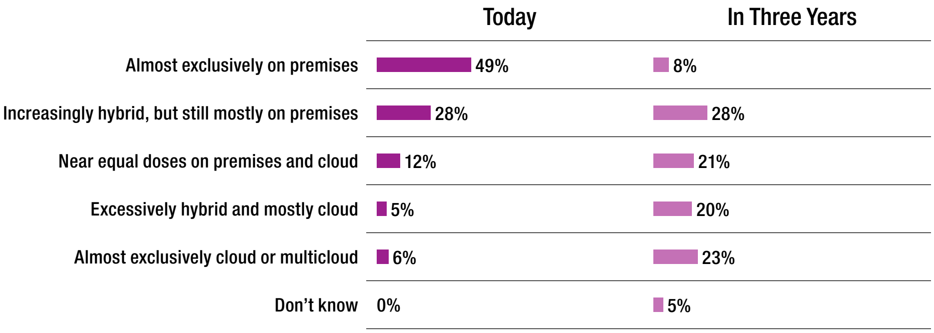


Figure 16. *Based on 99 respondents who have experience with DM for AA.*

Data volumes and their physical locations aside, our survey also asked about the origins of data for AA: Which of the following best describes the origins of data managed for advanced analytics? (See Figure 17.)[6] The mixture of source types is clearly shifting, which results in less data from traditional enterprise sources (from 61% to 45%) and more of an even mixture of modern and traditional data. Modern data has a faster growth rate, more diverse data structures, and tends toward cloud storage, so managers need to update their DM for AA and data platform strategies accordingly.

Which of the following best describes the origins of data managed for advanced analytics?

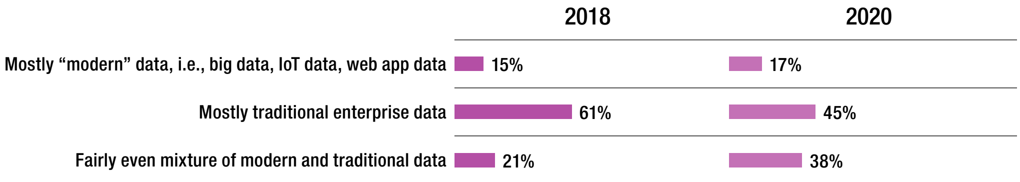


Figure 17. *Based on 99 respondents in 2020, 67 respondents in 2018.*

[6] A similar question was asked by the survey for the 2018 *TDWI Best Practices Report: Multiplatform Data Architectures* (See Figure 12 in that report.). The 2018 report is the source for the 2018 percentages included in this report's Figure 17. The 2018 data included 3% responding "Don't know," not shown.

Deploying AA Applications Leads to Complex, Hybrid Architectures

As we saw in the discussion of Figure 14, relational data and database management systems are now common in cloud and hybrid environments, which drives up the diversity of enterprise platform types. In a related trend, big data and other sources of modern data generate nonrelational data of diverse structures or no structure. For example, consider the proprietary and constantly evolving record structures generated by web applications and IoT sensors. Furthermore, unstructured data is on the increase from traditional enterprise applications (e.g., customer conversations captured by call center apps, the claims process in insurance, doctors' comments in medical records) and modern web apps (social media, e-commerce, containers, B2B data exchange).

As you embrace cloud and new data, rethink how you load-balance storage and processing for analytics data in hybrid architectures.

The result of all this is extremely diverse data. On the one hand, diversity is a blessing because it fuels rich correlations and modeling in AA. On the other hand, it is also a curse because diverse data structures, sources, and latencies further lengthen the list of data requirements for AA. A tried-and-true response by DM professionals is to satisfy diverse data requirements by deploying and optimizing a data platform for each group of requirements. This, in turn, jacks up the number of data platforms that must be acquired and supported. It also dramatically increases the complexity of data architectures, making them difficult to optimize and maintain over time.

Users find it increasingly difficult to provision data properly for analytics (76% in Figure 18). This is no wonder, given the exploding diversity of data. Traditional data platforms are not going away because they still manage large volumes of highly valuable data and they fit into business processes quite ably. Users are maintaining older platforms while adding new ones to address new data requirements as well as to scale at a reasonable cost. This is driving cloud adoption, and it explains how hybrid data architectures evolve.

As your organization adopts more forms of advanced analytics, do you find it increasingly difficult to provision data properly for analytics?

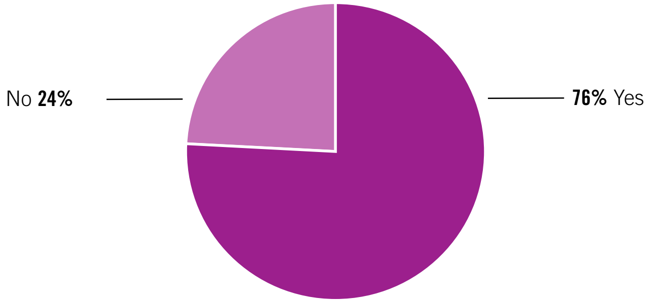


Figure 18. *Based on 155 respondents.*

Complexity is a problem for any IT system, but it is rarely a showstopper. For example, for decades DM professionals have been dealing with heavily distributed data strewn across multiplatform data architectures, as seen in the hybrid data architectures of modern data warehouses, analytics programs, multichannel marketing, and the digital supply chain. This kind of complexity is now the norm for DM for AA and other data-driven practices because it gives technical users many options for selecting the right platform for a specific form of AA.

Upgrade your data management architecture. Adopting cloud affects systems architectures for DM. Adapting to change is a challenge, yet it is also an opportunity to fix some of the mistakes of the past. For example, too many DM solutions evolve into a plague of point-to-point interfaces and integrations, resulting in a convoluted hairball that is hard to optimize, control, and maintain. Many organizations fix this common data management design problem by restructuring the hairball into a hub-and-spoke architecture or a data hub with controllable spokes. With the right tools, users can orchestrate data flowing through the hub to control access (for security and governance), improve data (for quality, modeling, and semantics), and make data accessible to a wider range of users (via self-service and publish/subscribe). Orchestration via a hub can apply to all data, including data flowing to and from clouds.[7]

DM for AA Workers and Managers

This report's survey asked respondents with DM for AA experience to enter the job titles of people who implement DM solutions for AA. (See Figure 19.)

Technical personnel for DM for AA projects are dominated by data engineers, architects, scientists, and analysts.

Data engineers and DataOps (20%). The job title "data engineer" and the team title "DataOps" both vary a bit, but they usually refer to people who perform the deep, back-end data work of data integration involving ETL, transformation logic, and the design of complex data flows and data orchestration. Just as often, they are cross-trained so they can build substantial solutions for data quality, metadata management, data modeling, data replication, synchronization, and more. Data engineers are the people who roll up their sleeves to integrate the data that feeds DM for AA and other high-end data-driven solutions.

Data architects (18%). These folks bring two disciplines together: One is data modeling, typically at the local level, to design tables, schema, dimensions, hierarchies, and time series. The other is true data architecture, which focuses on the "big picture" of how substantial numbers of models, data sets, databases, data structures, and containers relate and integrate to form a large-scale structure such as a snowflake schema, a hub-and-spoke architecture, or a distributed hybrid environment. Data architects design databases, warehouses, lakes, and other persistent-data structures where the data needed by AA is stored.

Data scientists (13%). These are the superheroes of the data management world. They are deeply proficient in all data management disciplines—from integration and quality to modeling and architecture. They are also proficient in all aspects of analytics development and programming in several computer languages. Furthermore, data scientists are famous for assembling massive, complex data sets and pulling them together into credible analytics prototypes in a short time frame. This breadth of skills makes the data scientist an excellent "bridge" person who understands both sides of DM for AA and can coordinate teams on both sides, especially in the early design phases of an analytics project.

Data analysts (13%). The data analyst is similar to a data scientist but more focused on solutions. The data scientist focuses on epiphanies and prototypes, whereas the data analyst builds practical analytics models and production-worthy solutions based on those epiphanies. DM for AA that is successful in production won't happen without diligent data analysts.

DM for AA is a group effort that needs business and tech managers, IT resources, and various business people.

Business and technical management (13%). As with any large and complex project or program, DM for AA needs a number of business managers and technical managers. Among these, chief officers (1.5%) and VPs (1.5%) get projects moving and keep them focused on the business goals of upper management, whereas directors (7%) and managers (3%) direct the quotidian work of DM for AA.

[7] For detailed discussions of hybrid data architectures and architectures for data integration, read the 2018 *TDWI Best Practices Report: Multiplatform Data Architectures* and the 2014 *TDWI Best Practices Report: Evolving Data Warehouse Architectures*, both online at tdwi.org/bpreports.

Miscellaneous IT personnel (9%). Systems analysts (3%) are the unsung heroes of large and complex IT solutions. Without them, DM for AA will lack the hardware, software servers, and system integration that it needs for success. Likewise, database administrators (3%) are critical to the setup, upgrades, optimization, and ongoing maintenance of the numerous DBMSs and other data platforms—both on premises and cloud—that the storage and persistence pieces of DM for AA demand. Other IT personnel (3%) contribute to DM for AA by providing technical support, network bandwidth, and other data center resources.

Miscellaneous business people (6%). These people define business requirements and provide other guidance for DM for AA projects as business data stewards, business analysts (not technical analysts), and business subject matter experts (SMEs).

Enter the job titles of people who contribute significantly to the design and implementation of data management specifically for advanced analytics:

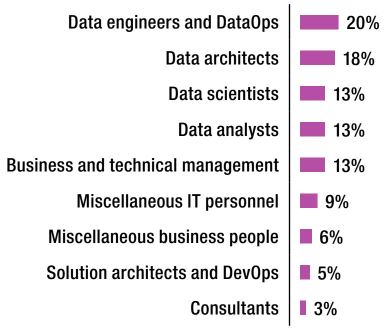


Figure 19. *Based on 152 responses from 99 respondents who have DM for AA experience; 1.5 responses per respondent on average.*

USER STORY HYBRID DATA ARCHITECTURE CAN INCLUDE A LEGACY DATA WAREHOUSE ON PREMISES AND A NEWLY DESIGNED ONE IN THE CLOUD.

“We all share a university-wide data warehouse,” said Sonj McCoy, a BI analyst at George Washington University (GWU). “It has a traditional design consisting of several data marts. Users go through an approval process—similar to data governance policies—to gain access to individual marts. This legacy warehouse continues to be valuable for most reporting applications.

“However, our legacy warehouse is fed by overnight batch ELT only. This means we are limited to latent reporting because data in the warehouse is always 24 hours behind. Also, the older design predates requirements for new analytics.

“To address those limitations, we also have a data warehouse on cloud, which is new for us. Some tools are on cloud, too, for integration and analytics. The cloud data warehouse supports real-time interfaces so we can get live, fresh data for time-sensitive reporting and analysis. The official GWU tool for building reports based on cloud data is a leading data visualization tool.

“We anticipate moving more data and functions to our cloud data warehouse. Yet, we will probably retain our legacy warehouse on premises as well because it still works fine with older reports and maintaining historical data.”

DM for AA Requirements per Analytics Use Case

DM Requirements for Set-Based Versus Algorithm-Based Analytics[8]

Forms of analytics divide into two broad categories, each with consistently recurring DM requirements.

Before we dive into the details of DM requirements for common forms of AA, we need to take a moment to establish two broad categories of analytics. TDWI distinguishes between the general approaches of “set-based analytics” (e.g., OLAP, ad hoc queries, SQL-based analytics) and “algorithmic analytics” (where algorithms parse and process data with little or no reference to data’s structure). One ramification for data requirements is that the first assumes that analytics data is structured (usually by the relational paradigm), whereas the second can handle data that is freeform in the extreme. Some discovery-oriented forms of analytics can work with either approach, as seen in data mining and machine learning. The point is that analytics tools tend to be one or the other—set-based or algorithm-based—and you will need to select tools and/or prepare analytics data accordingly.

Set-Based Analytics

This type of analytics assumes data is structured in sets (e.g., tables, dimensions, hierarchies, time series) with metadata or similar data semantics. Set-based analytics almost always operates on data managed in a relational DBMS, as seen in the common use cases described below.

- **Ad hoc queries and other SQL-based analytics** require SQL support that has optimized query performance. The data being operated on has traditionally been managed in relational DBMSs, though columnar DBMSs have become preferred because of their speed and storage optimization with columnar data. After all, most analytics queries are column-oriented. Finally, SQL over standard interfaces continues to be a highly effective data transport medium for the whole decision-making technology stack, from data integration and federation to advanced forms of analytics such as self-service and even machine learning.
- **Self-service analytics is mostly query-driven**, which is why it assumes the relational paradigm, SQL support, and some kind of metadata. Relational DBMSs and SQL tools are common enablers of self-service, though some users prefer to fulfill these DM requirements with Hadoop or a NoSQL DBMS. Self-service is a typical end-user requirement for data lakes, and data lakes originated on Hadoop, which explains why self-service and Hadoop sometimes go together.
- **Dashboards and performance management** rely on DM to provision structured data, usually in tabular form. Sometimes these practices require hierarchies when measures roll up into metrics, which roll up into calculated metrics, which roll up into KPIs.
- **Online analytical processing (OLAP) is all about dimensional data**, but there are many ways to model and store dimensions. Dedicated OLAP DBMSs and engines have gone away because today’s preference is to model dimensions with star schema or snowflake schema, which is easily done with commonplace modeling tools and relational DBMSs. With the speed and scalability of today’s data platforms, dimensions are increasingly instantiated on the fly instead of pre-stored.
- **Standard business reports depend on relational data**, typically provisioned from a traditional data warehouse. However, some modern forms of operational reporting (e.g., those that count website clicks, parse enterprise server logs, and count entities in unstructured sources such as IoT) rely more on data mining or natural language processing than the relational paradigm.

[8] Some of the material in this section of the report comes from the 2019 *TDWI Checklist Report: Data Management Requirements for Data Science, ML, and AI.* Available online at tdwi.org/checklists.

Algorithm-Based Analytics

Data structure is not an assumption for this type of analytics, and therefore it doesn't matter as much as with set-based analytics—in either sources or targets. Most recently created analytics algorithms can parse any data structure or deduce structures from seemingly unstructured sources (such as human speech).

- **A strong determinant of DM requirements for algorithmic analytics is where the data is stored**, more so than data structure. This is because many algorithms (or the tools that control them) prefer specific types of files, documents, or containers while others prefer to process data via "in-database analytics," but only for one or a short list of data platforms. Beware that older AA toolsets tend to be buckets of black-box algorithms, each with a specific data requirement, such as a recurring record structure in a large flat file.
- **Analytics algorithms are difficult to generalize** because of the numerous variances among them. Therefore, you must look at each analytics tool (and each of its individual algorithms) to determine the DM requirements for specific use cases. If you prefer to program your own algorithms, you will need to consider all these characteristics as you design.
- **The outputs from analytics algorithms vary considerably, too**. This differs from the predictable relational results of set-based analytics and can lead to problems. For example, some older data mining and clustering algorithms output a neural net; unfortunately, neural nets are usually problematic to store and re-instantiate from storage, proprietary to the originating tool, and therefore nearly impossible to share at a time when sharing analytics outputs is all the rage.
- **Analytics users today look for tools that produce a practical and sharable product.** For example, many algorithms for data mining and text analytics output structured records, which are easily used with relational DMBSs and SQL-based tools. As analytics tools consolidate into unified data and analytics platforms (UDAPs), the platforms provide sharing functions for data sets, visualizations, and analyses. Always consider how and by whom the analytics output will be used before making tool or design decisions.

DM Requirements for Data Mining and Other Discovery Analytics

Data mining algorithms vary, but most assume a large data set of raw source data, unaltered after extraction from source systems. Algorithms optimized for detailed source data are, by nature, tolerant of data with little or no structure or metadata. There are exceptions, for example, some mining tools need a proprietary schema or a flat file (not a DMBS).

For the broadest discovery experience, feed data mining vast volumes of detailed source data.

Data mining's focus on raw, detailed source data without metadata has ramifications for data management. In general, preparing data for data mining is largely about collecting data, not improving it. Hence, there is little or no remodeling, data transformation, data quality processing, or metadata development. In other words, data quality and transformation functions are not high priorities—as they are in DM for reporting and data warehousing.

In fact, improving data's quality and structure can be detrimental to data mining and other forms of discovery analytics (such as self-service data exploration via queries). Anomalies and outliers may reveal a new customer segment or a bad business practice. Nonstandard data can reveal fraud. Strip these out, and you've lost data nuggets that are valuable for the data discovery process.

Finally, data mining is one of many AA techniques that now relies heavily on data lakes. We could argue that the data lake evolved to serve data mining and discovery analytics in general. Most data lakes ingest data quickly to make data available for analytics and reporting as soon as possible. Storing incoming data in its arrival state speeds up and simplifies the ingestion process. More to the point, the initial state of data (i.e., raw and unaltered, detailed source data) is exactly what discovery analytics needs. Furthermore, respecting and maintaining the arrival state in data lake storage also means that original source data is available for repurposing as unforeseeable analytics projects arise in the future.

DM Requirements for Natural Language Processing (NLP)

Human speech and other unstructured data have potential business value.

There are different kinds of unstructured data, and NLP focuses on unstructured data sources that contain human language. Sources include the text fields of operational applications, flat files dumped from social media, and business productivity documents. Human language is captured as text in various industries: the claims process in insurance, caregiver notes in healthcare, call center and tech support applications in many industries, and the specialized documents exchanged among firms in business-to-business partnerships (such as in manufacturing, logistics, and retail). All these are valuable sources for analytics and business operations—if you have NLP tools to process them.

Text mining, text analytics, and other forms of NLP have evolved to process human language and other unstructured sources. Note that these are quite different from the DM techniques applied to traditional sources of analytics data. For an NLP-driven solution, source data often arrives in a file or document (which may be standard or proprietary), and the NLP tool must be able to parse these optimally. Other data comes from text fields in databases, which are commonly dumped into flat files, not accessed via queries. Metadata for these sources is minimal or nonexistent; however, some file and documents types (especially EDI, XML, and JSON) may have headers that are equivalent to metadata. The most common data type is "text," which is almost meaningless; the NLP tool must add value by giving output data structure and metadata labels.

DM must be extended with NLP to realize the analytics potential of unstructured text.

The "secret sauce" in most NLP-driven solutions is "entity extraction," sometimes referred to as "ETL for text." Here's how it works. An NLP tool makes multiple passes through unstructured data. In an early pass, it locates text about business entities of interest, typically the names of customers, people, firms, products, and locations (names of states, towns, streets, stores, neighborhoods), plus dates and times. On subsequent passes, the tool identifies facts around each entity instance, then packages all these data points into a fact record per entity instance. Ironically, NLP parses so-called unstructured data sources to yield a highly structured analytics output.

Once NLP's fact records go into a fact table or similar structure, the output of NLP analytics is easily consumed by a wide range of tools and users and stored in data warehouses or lakes. Hence, NLP is an effective feeder technology that provides valuable data for other AA methods, commonly sentiment analysis, statistical analysis, data mining, and self-service, as well as non-AA use cases in reporting and operations.

DM Requirements for Real-Time Analytics

Real-time AA isn't possible without real-time DM.

For decades, reporting and analytics tools and best practices have been moving closer and closer to high performance in real time (sub-second response time), near time (minutes or hours), or right time (a response frame determined by the practical needs of a business process). Overnight batch processing is still the norm for refreshing most business reports and production analytics, yet this is increasingly complemented by real-time DM and real-time analytics in support of time-sensitive business processes.

Note that real-time data and analytics are only appropriate to business processes that can benefit from fresh information and can actually react in close to real time. For example, true real-time analytics is well established in utility firms, stock trading, industry exchanges, shopper recommendations in e-commerce, and many manifestations of surveillance or business activity monitoring. Near-time analytics (or hourly operational reporting and dashboards) can enlighten shop floor management in manufacturing, hospital bed allocation in healthcare, and optimization of daily operations in logistics and supply-chain-oriented firms.

Real-time analytics assumes real-time data management. On the inbound side, DM infrastructure must handle real-time data pushed to it via streams, feeds, IoT, and a wide range of application interfaces. For some sources (typically operational applications that generate time-sensitive data), DM must extract data periodically during the business day, process it, and deliver it to AA tools. On the outbound side, DM infrastructure regularly accepts output from AA tools and in real time integrates them with other tools or data warehouses and data lakes.

These deployment designs require tight interoperability between DM infrastructure and AA tools as well as extreme high performance from both. They also require functionality outside the usual DM toolkit, such as event processing or complex event processing, perhaps application integration or some form of middleware. In many use cases, microbatching will suffice for near-time or right-time requirements.

Modern Data Semantics Requirements for DM for AA

There are now several established forms of data semantics, namely metadata management and multiple forms of metadata (e.g., technical, business, and operational metadata), as well as emerging semantics for business glossaries, data profiling, and data cataloging. Because modern users want to query, browse, and search semantic descriptions of data (which leads to accessing the data), a modern semantics facility must support multiple forms of indexing, including keyword search indices. Sophisticated users (such as those performing DM for AA) are using all these approaches to data semantics, often in a single project.

Today's requirements demand multiple semantic approaches, from three forms of metadata to glossaries and catalogs.

The long list of approaches comes together under the umbrella term *data semantics*, and these descriptions of data are managed by a *modern semantics facility*. Note that most semantics facilities are actually metadata management or data virtualization platforms that have been modernized and expanded to handle far more than metadata and virtualization.

The modern semantics facility provides many options for DM and AA. When a modern semantics facility is centralized and shared, it presents a comprehensive inventory of data for all the platforms involved in DM for AA, whereas traditional semantics rarely reaches beyond metadata for a single platform. A semantics facility enables the creation of custom views of distributed data, such as business metadata or glossaries for business users.

Ideally, the same facility can also enable sophisticated data virtualization (DV) and high-value applications of DV, such as the logical data warehouse or logical data lake. Finally, note that semantics-driven views or virtual applications can impose architectural unity upon the siloed chaos of the hybrid data architectures (HDAs) typical of AA nowadays, without the risk, cost, and distraction of time-consuming data migration and consolidation projects. That's why the modern semantics facility is becoming one of the leading tools for the unification of HDAs and other complex distributed data architectures, especially in relation to AA use cases.[9]

Metadata management across new platforms. For example, modern metadata management tools are now appearing as SaaS platforms. The benefits of SaaS and cloud-based tools apply to metadata management, namely minimal tool setup, tool maintenance, and capital investment,

[8] The modern semantic facility is also a critical success factor for data hubs, as explained in the 2018 *TDWI Checklist Report: The Modern Data Hub.* Available online at tdwi.org/checklists.

with short time to use and elastic scale in production. As a completely different example, metadata tools must interface with data management functions on new platforms, such as SaaS operational apps, cloud-based systems and storage, Hadoop, and other open source. Finally, given the multiplatform, hybrid data environments becoming popular today, it sometimes makes sense to deploy a hybrid metadata repository that stores metadata on diverse platforms, although its interface makes distributed metadata look like a single source.

Data management and analytics professionals turn to technologies designed for cross-platform data operations in cloud and hybrid architectures, such as data virtualization, query federation, integration hubs, data flows, and data replication. However, they also need semantics that draw a "big picture," as done by enterprise data catalogs, business glossaries, and modern approaches to metadata management. These big-picture functions contribute to multiple contexts, including AA development, DM for AA, run-time deployment, data governance, and self-service data access.

Data Virtualization as a DM Strategy for AA

Data virtualization is a form of data integration that provides abstraction and services layers as a virtual complement to physical integration. DV is an appropriate DM strategy when AA applications demand very fresh data or when AA data is distributed across multiple data platforms in a hybrid data architecture.

Data virtualization enables compelling use cases for DM and AA.

- Many virtualized data services can operate in real time (or close to it) to instantiate fresh data that is time sensitive for business processes or updated repeatedly during the business day. Hence, DV can enable real-time analytics and embed analytics into operational applications.
- Virtual views are often designed to be business-friendly and can simplify access to analytics data as required for self-service data prep, exploration, and visualization.
- DV reduces data replication and relocation, which reduces network loads, storage consumption, and the redundancy of analytics data (which can skew analytics outcomes).
- Data may be migrated through data virtualization's abstraction layer for fast prototyping and testing of AA solutions. DV infrastructure also facilitates migrations and consolidations of data into target systems on premises or in the cloud.
- DV techniques create data interfaces and communication channels among the many platforms of an analytics ecosystem, which in turn unifies the large-scale architecture of that ecosystem.

DM and Other Requirements for Self-Service Analytics

Self-service practices give new classes of end users productive and autonomous analytics.

For years, TDWI has seen self-service data practices inching ever upward in popularity across all industries and organizational sizes. These practices are often performed in a series of tasks from self-service data browsing and querying to data visualization and collaboration. Usually, it is business people—with enough data and tool skills to form queries and create their own visualizations—who are demanding self-service functionality. Self-service DM and analytics enables business users to work autonomously instead of voicing their requirements to IT or a data team, then waiting weeks for the data they need. Furthermore, self-service tools are famous for high ease of use, which helps the end user be quick and productive. TDWI sees self-service used in diverse scenarios, including offloading BI work from technical users to business ones, supporting departmental BI without a data warehouse, and enabling data discovery and data prep for data sets on a data lake or warehouse, whether on premises or in the cloud.

Although self-service data and tools are almost always deployed for mildly technical business users, some technical users find them useful, too. When a data scientist or data analyst receives a new assignment—for example, to quantify the latest form of customer churn—they can get a quick and dirty read of data before deciding which direction to go with more advanced tooling.

To gauge the urgency of self-service tools and the need for end-user autonomy, our survey asked: How important is it to enable end users to manage their own data sets without IT support? (See Figure 20.)

The vast majority of respondents recognize the need for end-user autonomy (95%). In other words, self-service is either very important (44%) or somewhat important (51%). Almost no one claimed it is not at all important (5%).

How important is it to enable end users to manage their own data sets without IT support?

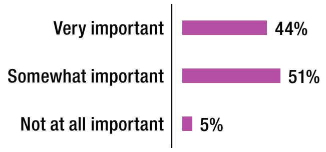


Figure 20. *Based on 99 respondents who have DM for AA experience.*

Self-Service Has Data Requirements, as with Any Analytics or Reporting Technique.

Self-service has many requirements for DM, data semantics, end-user tools, and a unified analytics experience.

The most important lesson learned about self-service analytics and other practices is that you cannot merely give end users access to data that was prepared in a mainstream way and expect them to succeed. In other words, self-service is like other disciplines in analytics and reporting—it has data requirements, and data management professionals must address them to ensure that self-service solutions are useful and valuable enough to be considered successful. Luckily, in recent years, DM and AA best practices have coalesced for preparing data, tools, and end users for successful self-service work, as described below.

Self-service end users need business-friendly tools with high ease of use for data access. Handing them a BI tool built for a developer will not suffice. Instead, organizations are trending toward the recent generation of data visualization tools (which do far more than just visualization). However, some organizations are deploying a unified data and analytics platform (UDAP), which is a collection of tightly integrated tools that addresses all practices within the self-service discipline.

Self-service end users need business-friendly data semantics. Technical metadata will not work. Instead, business end users need business metadata or an equivalent approach to data semantics, as seen in the business glossary or data catalog. A common use of data virtualization is to create a single, business-friendly virtual view of far-flung analytics data, which does double duty as a single point of entry for security, governance, curation, and standards.

Self-service end users need ad hoc queries on steroids. After all, every new project on a self-service platform begins with data browsing and ad hoc querying. The end user then iteratively revises an ad hoc query and applies data prep functions until the result set contains exactly the information the user wants in a form that is appropriate to the next steps in self-service visualization or analytics.

Self-service end users need data prep, not ETL and a DW. Let's be honest: most visualization and analyses coming out of a self-service platform are simple, involving straightforward data transformations and not requiring a data warehouse for elaborately remodeled historical data. On the flip side, data prep has advanced impressively in recent years, making it more than capable of sophisticated data set design but with high ease of use.

Self-service end users need quality data in usable structures. Classic data quality functions—such as profiling and standardization—make queries easier to create and faster when they run. Data deduplication and matching help the end user cope with the redundant data that is typical with data lakes and other large repositories where self-service is applied. Ideally, data quality functions should be built into the data prep tooling of a self-service platform so quality is addressed on the fly as each new data set is created.

However, do not take data quality functions too far. Be content with "just enough" structure and light data standardization. Otherwise, you may strip out the anomalies, outliers, and nonstandard data that self-service data exploration and discovery-oriented algorithmic analytics depend on to spot potential fraud, new customer segments, and security incursions.

Self-service end users need all of the above interoperating in a unified analytics process. The holy grail of self-service is to perform it as a sequence of related steps—in a unified analytics process—seamlessly moving through browsing, ad hoc queries, data prep, visualization, publication, and collaboration. For the highest satisfaction among end users, DM and AA professionals should endeavor to provide such a unified user experience.

However, half of our survey respondents said they are "concerned about getting visualization tools and data sets to play well together." TDWI has encountered many organizations that successfully integrate data visualization tools and modern data platforms. In fact, this combination of DM and AA tooling is a de facto standard for self-service. On the horizon, however, several vendors are working on the UDAP mentioned earlier. Only time will tell, but the UDAP has the potential to become the preferred platform for self-service and similar analytics or reporting.

DM Requirements for Machine Learning

We conclude this report with an in-depth examination of the fastest growing discipline in advanced analytics, namely machine learning (ML). In terms of DM for AA, ML is an extreme and fascinating case because its complex life cycle of discovery, prototyping, predictive modeling, production, and maintenance involves at least five sets of data requirements. All these must be respected and satisfied to ensure high-quality modeling and predictive analytics.[10]

Machine learning is the hottest form of AA today and will continue to be so.

New advances in artificial intelligence are largely driven by machine learning. The fields of science and computing have been on an active quest for artificial intelligence (AI) for about 75 years, starting during World War II, when intelligent machines were first built for code breaking, calculating artillery trajectories, and predicting weather. Today, breakthroughs in AI are finally delivering on their promises to everyday business users, largely driven by advancements in ML. Whereas older approaches to AI were largely top-down and based on algorithmic logic, newer approaches based on ML are bottom-up in that they study data to identify patterns, relationships, correlations, outcomes, and other inferences. These data-driven discoveries, in turn, are incorporated into predictive models, which make production AI systems practical for predicting and remediating customer churn, fraud detection, machinery maintenance, efficiency improvements, software-driven automation for a wide range of organizational processes, and many other use cases.

[10] The material in this section of the report is drawn from the 2019 *TDWI Checklist Report: Data Management Requirements for Data Science, ML, and AI.* Available online at tdwi.org/checklists.

Tools and automation for machine learning are improving rapidly. Recent advances in AI and ML are impressive. Even more impressive is the fact that ML is evolving to require less initial human intervention and provide greater automation. For example, analytics tools for ML and AI are progressively more capable of parsing data, generating predictive models, putting models into production, and maintaining models over time with little or no guidance from developers. This end-to-end development process is often called automated machine learning (AutoML). In other words, data science is being used to automate the creation of new ML-driven AI models and algorithms, resulting in more solutions and simpler AI designs, more accurate and iterative models, created and deployed faster, by both ML experts and non-experts.

Similarly, production AI/ML algorithms and models (either standalone or embedded in an application or service) are becoming more capable of making decisions and taking action automatically. For example, it's already common in the real world for ML-driven predictive algorithms (embedded in an e-commerce application) to monitor a website visitor's behavior and then recommend an appropriate product or service.

ML and AutoML are only as good as the data fed to them. Here's the catch: the newfound abilities of ML and AutoML depend heavily on getting the right data at the right time to the correct models. Preparing data for ML is complicated by the fact that the ML life cycle has multiple stages, and each stage has slightly different data requirements that shift as the life cycle moves from discovery through development into production and beyond into maintenance and upgrade stages. The success of each ML life cycle stage depends on getting just the right data in the right condition onto data platforms that are conducive to data analytics. Even so, organizations with experienced teams and modern toolsets for data management have proved that they can satisfy the complex data requirements of machine learning. The next section of this report will describe ML's life cycle stages and their unique data requirements.

As with most AA forms, the predictive models output by ML are influenced by the data input for their design.

ML and AI are important because they can transform a business. To be competitive, agile, and growth-oriented, business managers today feel they need a broader range of advanced analytics based on statistics, mining, graph, natural language processing, and various predictive approaches. AI and ML certainly fit this trend by refining the vast stores of data collected by every business. However, modern business management also needs to take analytics out of the back office and push it into the front line, typically as part of customer interactions and other operational tasks. ML-driven AI embedded in operational applications is an effective way to realign advanced analytics with modern business processes.

In fact, many businesses today are already using AI and ML for these use cases. For example, the 2019 TDWI survey on AI and ML asked: "What kind of AI technologies do you currently use?" A whopping 92% reported using machine learning. The survey also asked "what AI use cases dominate?" Eighty-five percent reported "building predictive models using tools such as ML."[11]

The Five Life Cycle Stages of Machine Learning and Their Data Requirements

Developing an ML-driven predictive analytics solution or embeddable feature is similar to other development projects in that its life cycle includes several stages in sequence. The difference with ML is that its life cycle includes five demanding stages, namely those for solution definition, development, deployment, production, and monitoring output. To make ML even more challenging, each life cycle stage has distinct data requirements, as illustrated in Figure 21 and discussed below:

Machine learning is extreme in that every modeling project must satisfy at least five sets of DM requirements.

1. **Define the problem and its solution.** As with most development projects, a machine learning solution should begin by defining the business problem (e.g., a new form of

[11] See the discussions of Figures 1 and 2 in the 2019 *TDWI Best Practices Report: Driving Digital Transformation Using AI and Machine Learning*, online at tdwi.org/bpreports.

customer churn) as well as a potential IT solution (an analytics model that can predict the new form of churn and recommend actions that prevent it). One way that business and technical people approach churn is by creating a data set that represents customers and their recent activities that resulted in churn—effectively a training data set. Other analytics solutions likewise begin with a data set.

- **Exploratory data.** Many user types depend on data exploration (sometimes called data discovery) to gain initial insights and to formulate a hypothesis in the earliest stage of solution development. So that exploration is not inhibited, data sets for exploration should be large (ideally terabytes), include lots of dimensions and details (as raw source data does), and contain information from many contexts (e.g., customer behaviors relative to sales, service, and purchasing in both historical and recent time frames).

2. **Prototype predictive models.** As noted earlier, whether a developer creates a predictive model manually or a smart AutoML tool generates it, you need learning data to work with. At this point, a rough data set will suffice for prototyping one or more analytics models, as well as to enable collaboration among the users that must review the prototypes. Once the desired model design is agreed upon, learning data will give way to training data.

   - **Learning data.** In many cases, the data set that results from data exploration and self-service tools can be improved via data prep to become suitable for learning. In other cases, data scientists, data analysts, and other data professionals may rely on ad hoc queries, data virtualization and federation, or ETL functions. Like most test data sets, the exact size of learning data is not critical, although bigger is better. It is more critical that the learning data contains broad information about the entities, activities, and processes being modeled.

3. **Select a model and tune its hyperparameters.** Today's rapid prototyping typically leads through multiple prototypes and/or multiple iterative versions of a prototype. This process drives toward the selection of a predictive model design that most closely matches the characteristics of the desired analytics solution. Once selected, the model's hyperparameters should be tuned for the solution's target prediction and scoring. Then it is time to create training data specifically for the selected predictive model and its parameters.

   - **Training data.** The final production version of the predictive model will be generated from (or otherwise based on) the training data. Therefore, to ensure that the production model is fully relevant to the solution's intended target predictions, training data (like learning data) should contain broad information about the entities, activities, and processes being modeled, but cleansed and targeted to the needs of the final model version and its parameters. To ensure an adequate sample for statistics, clustering, and networks is generated for the model, training data is typically much larger than learning data. Data quality is very important to training data, in that it should be cleansed of outliers and nonstandard data that would skew the model's training and design. When assembling training data, be sure that the same data is not loaded redundantly, which can also skew models. If you will be using an AutoML tool, be sure that training data complies with the formatting and input data requirements of the tool; for example, some tools demand file-based data, specific schema, or a single table.

4. **Deploy the solution with the predictive model.** This varies widely; some predictive functions are algorithms or services embedded in larger applications while others are full-blown, standalone analytics applications.

- **Input production data.** There are at least two types of production data. First, there is data that is periodically fed into the model for scoring; this is usually some kind of operational data that may be integrated straight from operational applications, through middleware, or after being persisted in a data warehouse, lake, or similar database. This data is relevant to other forms of analytics and so should be captured (usually in a data warehouse) for reuse and later study.
- **Output production data.** The second type of production data is the output of the model, typically scores and codes denoting the probability of the entity state or action being predicted. The data output should be routed or stored for an appropriate time frame depending on who will use it and how.

5. **Monitor model output.** Over time, the behavior and characteristics of the entities present in the original training data will change (e.g., this is common with customers and partners). Likewise, input production data may change its schema, quality, or frequency of generation (due to application upgrades, changes in end-user usage, new applications coming online, or the addition of new data sources). Such influences can cause a production data model to "drift" from its original concept, tuning, or predictive accuracy. Hence, it is best to regularly monitor model inputs and outputs, then periodically retrain the model via machine learning to ensure its continuous improvement, relevance, accuracy, and adaptation to change via learning.
   - **Retraining data.** When drift is minimal, consider simply reusing the original training data to retrain the model, but with more recent data appended that represents changes in the entities modeled. When drift is more dramatic—or it is time for a major update to create new functionality—developers may start over with a new set of learning data, then use ML to design or generate a new model and create fresh training data. Obviously, the new version of the predictive model will need to go through side-by-side testing before being redeployed to determine whether it is actually more accurate.

The Five Life Cycle Stages of Machine Learning and their Data Requirements.

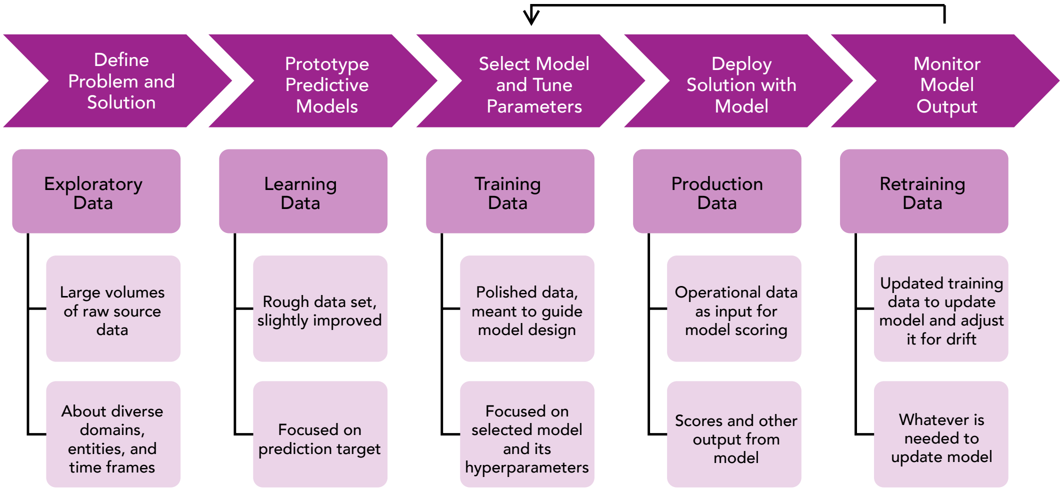

Define Problem and Solution
Prototype Predictive Models
Select Model and Tune Parameters
Deploy Solution with Model
Monitor Model Output
Exploratory Data
Large volumes of raw source data
About diverse domains, entities, and time frames
Learning Data
Rough data set, slightly improved
Focused on prediction target
Training Data
Polished data, meant to guide model design
Focused on selected model and its hyperparameters
Production Data
Operational data as input for model scoring
Scores and other output from model
Retraining Data
Updated training data to update model and adjust it for drift
Whatever is needed to update model

Figure 21. *Source: TDWI 2019.*[12]

[12] See Figure 1 in the 2019 *TDWI Checklist Report: Data Management Requirements for Data Science, ML, and AI.* Available online at tdwi.org/checklists

USER STORY THE DATA ECOSYSTEM IS FIRST AND FOREMOST. ANALYTICS IS BUILT ON THAT.

"My mandate is to create an enterprise-scope view of the firm via analytics," said Greg Nelson, former VP, analytics and strategy, at healthcare provider Vidant Health and author of *The Analytics Lifecycle Toolkit*, published by Wiley and Sons. "As a critical first step, my team and I have focused on getting data and its management under control. Our first step in that was to get a clear picture of all data going in and out, as well as within the organization, and catalog its use and provenance. We ended up cataloging hundreds of data feeds and their quality requirements across fifty-four enterprise systems. To give that cataloging effort fuel, we designed a data governance program and clarified our data strategy.

"Concurrent with improvements to our data ecosystem, we were able to launch over two dozen system-level dashboards. These dashboards, along with those in development, gave us much needed actionable data around key strategic measures, such as employee turnover, financial performance, and supply chain. We're working on real-time operational models so we can see and forecast capacity throughout the system and quickly identify gaps that may require real-time resolution. We're also putting a lot of energy toward machine learning to get predictive algorithms and models for employee churn in nursing, patient capacity, surgery demand, and unauthorized data access. Forecasting COVID-19 patients is obviously a priority for predictive analytics today."

Top Twelve Priorities of Data Management for Advanced Analytics

In closing, we summarize this report by distilling from it the top priorities for achieving successful DM for AA. We also reflect on why these priorities are important. Think of the priorities as recommendations, requirements, or rules that can guide user organizations through a successful DM for AA program.

The first four priorities are the guiding principles of DM for AA.

1. **Realize that advanced analytics is not one thing.** Instead, AA contains many approaches, and each has its own purpose, value proposition, use cases, and enabling technologies. Knowing the characteristics of each is fundamental to making good decisions about which to use when.

2. **Always remember that each approach to AA has distinct requirements for DM.** Fully satisfying DM requirements leads to a successful AA solution. Data and analytics professionals who ignore the requirements hamstring their solutions and risk failure.

3. **Bring the disciplines of AA and DM closer together.** This ensures that AA solutions get data from the most appropriate sources, containing rich information about business entities of interest, in the best schema for the AA tools being used, with an acceptable level of quality, delivered at the right time, through an optimal interface.

4. **Tailor your DM practices and tool usage to the needs of individual analytics solutions.** In other words, you cannot perform DM for AA in a single way and expect all implementations of advanced analytics to yield equally useful and accurate outcomes.

Solution designs and tool selection hinge on the pairing of DM options with specific AA approaches.

5. **Understand AA/DM pairings. Design DM for AA infrastructure with them in mind.** For example, the large volumes of human speech captured in text files that are required for natural language processing differ radically from the lightly standardized tabular data required for self-service data practices. As another example, algorithms for data mining, statistics, and machine learning work well with unstructured or inconsistently structured data

of poor quality and with no metadata, whereas data warehouse analytics based on time series, hierarchies, or dimensions demands ruthlessly structured, cleansed, and documented data.

6. **Beware that some AA approaches have multiple sets of DM requirements.** For example, each of the five life cycle stages of machine learning has its own DM requirements. Similarly, in a unified self-service analytics process, each step has slightly different DM needs, yet DM artifacts (metadata, data sets, security roles, GUI) must be shared across all steps. In addition, mature users regularly deploy multiple AA solutions, even within a single project, and you must satisfy the DM requirements of all approaches.

7. **Let your knowledge of AA/DM pairings guide the selection of data platforms and tools.** For analytics approaches that demand massive data volumes (e.g., mining, clustering, statistics), users tend to deploy Hadoop or a cloud-based DBMS for their analytics data. Some analytics run best "in database," so you should acquire data platforms that support the kind of in-place processing you need. For real-time analytics, you will need tools for real-time data ingestion. To succeed with self-service analytics, you need solid business metadata and possibly a data catalog.

8. **Embrace DM for AA for its benefits.** According to the survey, these include improvements to operations, analytics outcomes, DM upgrades, and real-time data and analytics.

9. **Create contingency plans for DM for AA's potential barriers.** Survey results say that problems may arise in data governance, architecture, skills, and DM infrastructure.

DM for AA offers many compelling benefits, but suffers from a few minor challenges.

10. **Remember that data integration is just as important as data platforms in your DM strategy for AA.** The continuing hype around new data platforms can distract us from other key technologies for DM for AA, in particular data integration, data quality, data virtualization, real-time event processing, and metadata and other modern data semantics.

11. **Support the DM and AA professionals who make DM for AA happen.** According to survey responses, the leaders are data engineers and DataOps (20%), data architects (18%), data scientists (13%), and data analysts (13%).

12. **Expect multiple types of cloud-based systems to be part of your DM for AA strategy.** The general benefits of cloud (mostly based on elasticity) apply amply to DM for AA, so you should seriously consider cloud data platforms and cloud-based tools for data integration and AA. Also, expect to redesign your data architectures periodically as they go hybrid due to cloud adoption. Finally, survey data suggests that the amount of analytics data managed on cloud platforms will quadruple within three years. Address this challenge via cloud elasticity because it is the modern way to ensure ever-increasing storage and processing capacity.

Cloud is quickly becoming the preferred compute platform for data-driven disciplines, including DM for AA.

denodo

www.denodo.com

Denodo is a leader in data virtualization providing agile, high-performance data integration, data abstraction, and real-time data services across the broadest range of enterprise, cloud, big data, and unstructured data sources at half the cost of traditional approaches. Denodo's customers across every major industry have gained significant business agility and ROI by enabling faster and easier access to unified business information for agile BI, big data analytics, web and cloud integration, single-view applications, and enterprise data services.

The Denodo Platform offers broad access to structured and unstructured data residing in enterprise, big data, and cloud sources, in both batch and real-time, exceeding the performance needs of data-intensive organizations for both analytical and operational use cases, delivered in a much shorter time frame than traditional data integration tools.

The Denodo Platform drives agility, faster time to market, and increased customer engagement by delivering a single view of the customer and operational efficiency from realtime business intelligence and self-serviceability.

For more information visit www.denodo.com, follow Denodo via Twitter @denodo, or contact us to request an evaluation copy at info@denodo.com.

research

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management 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 data management 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