

Data Management Requirements for Data Science, ML, and AI

Getting the Right Data in the Right Format
for Data Science and Analytics



By Philip Russom

Sponsored by:



JANUARY 2020

TDWI CHECKLIST REPORT

Data Management Requirements for Data Science, ML, and AI

Getting the Right Data in the Right Format for Data Science and Analytics

By Philip Russom



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

TABLE OF CONTENTS

- 2 **FOREWORD**
- 4 **NUMBER ONE**
Understand the ML life cycle and how data requirements vary across life cycle stages
- 7 **NUMBER TWO**
Consider a modern cloud data warehouse as your primary platform for ML data
- 9 **NUMBER THREE**
Consider a data lake as a secondary platform that satisfies multiple ML data requirements
- 11 **NUMBER FOUR**
Note that data integration is a critical success factor for ML
- 13 **NUMBER FIVE**
Remember that analytics tools have data requirements, too
- 14 **NUMBER SIX**
Select your first steps in ML-driven AI and predictive analytics based on business need and sensible project planning
- 16 **ABOUT OUR SPONSOR**
- 16 **ABOUT TDWI RESEARCH**
- 16 **ABOUT THE AUTHOR**
- 16 **ABOUT TDWI CHECKLIST REPORTS**

© 2020 by TDWI, a division of 1105 Media, Inc. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. Email requests or feedback to info@tdwi.org.

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.

FOREWORD

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 machine learning (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 fraud detection, machinery maintenance, efficiency improvements, predicting and remediating customer churn, software-driven automation for a wide range of organizational processes, and many other use cases.

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, simpler AI designs, and more accurate and iterative models, created and deployed faster, by both ML experts and nonexperts.

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 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 stage depends on getting just the right data in the right condition onto data platforms that are conducive to data analytics.¹ 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.

¹ In this report, the umbrella term "analytics" refers to advanced forms of analytics, based on statistics, mining, clustering, graph, neural networks, machine learning, artificial intelligence, and so on. Note that the term does not encompass forms of business intelligence (BI), such as reporting, dashboards, and OLAP.

FOREWORD CONTINUED

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.

Furthermore, business people are tired of offline, after-the-fact analytics or reporting that forces them to drive while “looking in the rear-view mirror” at yesterday’s corporate performance. They need to look through the windshield to see and predict customer churn, competitive threats, revenue shortfalls, and fraud in real time (or close to it) while there is still time to react and prevent, not merely detect. Predictive technologies, such as AI and ML, are key to modern preemptive business tactics.

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?” and 85% reported “building predictive models using tools for ML.”²

This TDWI Checklist report drills into the business and technical requirements for preparing data for machine learning and related practices in artificial intelligence and predictive analytics. The goal is to help user organizations understand the new data requirements for AI/ML-driven analytics as well as how to satisfy those requirements with platforms, tools, and best practices in data science.

² 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.

1

UNDERSTAND THE ML LIFE CYCLE AND HOW DATA REQUIREMENTS VARY ACROSS LIFE CYCLE STAGES

Developing an ML-driven artificial intelligence solution or embeddable feature is similar to other development projects in that its life cycle includes several stages in sequence, namely those for solution definition, development, deployment, production, and updates. The difference is that the machine learning life cycle has distinct requirements for each stage, as illustrated in Figure 1 and discussed below.

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 customer churn) as well as a potential IT solution (e.g., an analytics model that can predict any 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 as not to inhibit exploration, data sets for exploration should be large (ideally terabytes) and should include many dimensions and details (as raw source data does) as well as information from many contexts (e.g., customer behaviors relative to sales, service, purchasing, both historic and recent).

PROTOTYPE PREDICTIVE MODELS. As noted earlier, whether a developer creates a predictive model manually or a smart AutoML tool generates it, you need so-called 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

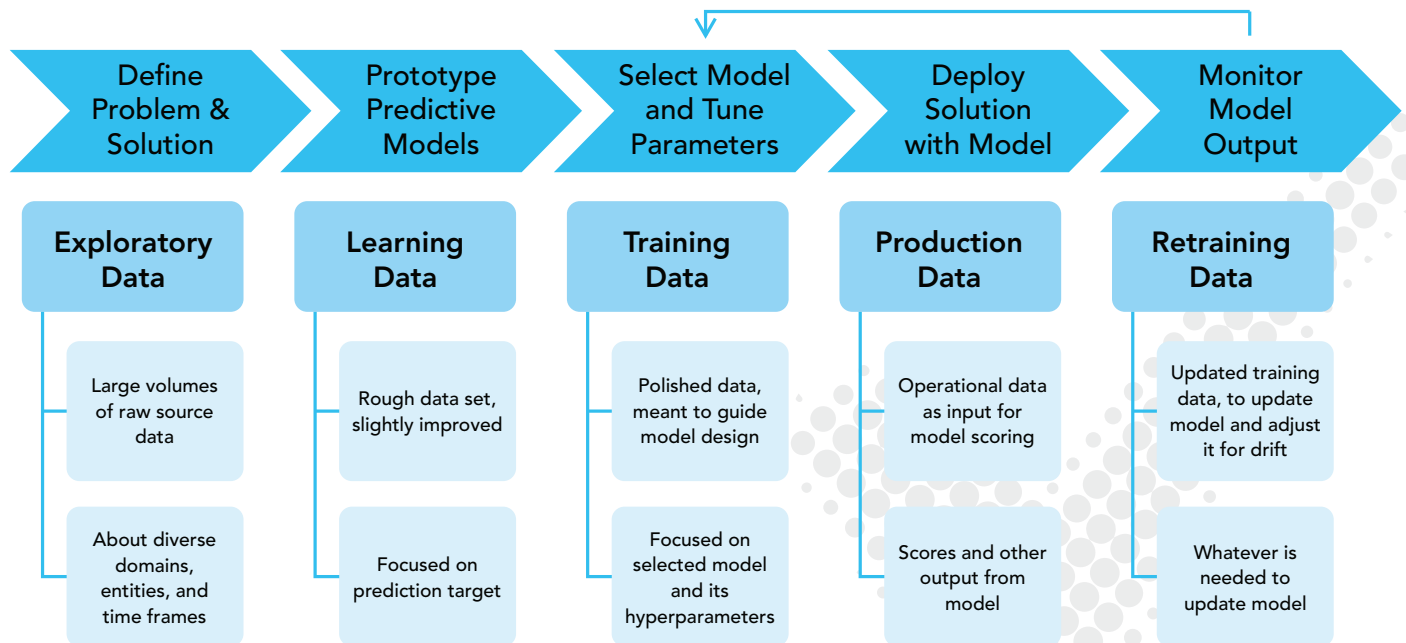


FIGURE 1. The life cycle stages of machine learning, with data requirements per stage.

UNDERSTAND THE ML LIFE CYCLE AND HOW DATA REQUIREMENTS VARY ACROSS LIFE CYCLE STAGES CONTINUED

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.

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. Furthermore, 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 assure that the production model is fully relevant to the solution's intended target predictions, training data (as with learning data) should contain broad information about the entities, activities, and processes being modeled but cleansed and more targeted to the needs of the final model version and its parameters. To assure an adequate sample for statistics, clustering, and networks 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 the training data complies with the formatting and input data requirements of the tool; for example, some tools demand file-based data, specific schema, a single table, and so on.

DEPLOY THE SOLUTION WITH THE PREDICTIVE MODEL. This varies considerably because 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, in an appropriate time frame, depending on who will use it and how.

UNDERSTAND THE ML LIFE CYCLE AND HOW DATA REQUIREMENTS VARY ACROSS LIFE CYCLE STAGES CONTINUED

MONITOR MODEL OUTPUT. Over time, the behavior and characteristics of the entities present in the original training data will change (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. It is best to regularly monitor model inputs and outputs, then periodically retrain the model via machine learning to assure 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.



2

CONSIDER A MODERN CLOUD DATA WAREHOUSE AS YOUR PRIMARY PLATFORM FOR ML DATA

As we just saw, machine learning is a data-intensive method for designing a predictive model, and its success demands large volumes of diverse data to be collected, persisted, transformed, and presented in many different ways. The question is: what type of data platform can handle the volume, diversity, processing, and relentless repurposing of data seen in machine learning?

A modern built-for-the-cloud data warehouse platform is an obvious choice because it can satisfy the data requirements of all life cycle stages in machine learning, artificial intelligence, and predictive application development.

For example, the general benefits of the cloud apply to data warehousing and analytics, namely:

- Unlimited and affordable scalability for the storage of massive volumes of structured and semistructured data
- The ability to analyze all the data quickly with minimal latency
- High-performance data access and in-place processing for analytics and data integration powered by the cloud's elastic resource allocation and decoupling of compute and storage resources
- Low cost of entry because the cloud eliminates capital expenditures for on-premises hardware and its system integration
- Low total cost of ownership due to usage-based pricing and data center outsourcing
- Minimal time required for system configuration, upgrades, administration, and capacity planning

- DataOps, data scientists, and analysts can focus on creating new analytics instead of moving data around

In addition, a modern cloud data warehouse specifically addresses many data requirements for machine learning:

NEW CLOUD DATABASES ARE RELATIONAL—AND MORE. They support SQL and the rest of the relational paradigm, which are well understood skills that can be efficient approaches to preparing data and scoring models for ML, AI, and predictive analytics.

NEW CLOUD DATABASES CAN EXECUTE POPULAR PROGRAMMING LANGUAGES. This is important because analytics algorithms are increasingly written in R, Python, Java, and Scala.

A MODERN CLOUD DATA WAREHOUSE WILL SUPPORT MANY PROCESSING ENGINES. This includes multiple push-down and in-database processing options (executing SQL, R, Python, or Scala in the database). Scalable in-situ processing is important because big data is too big to move around for ML and other analytics. Plus, the trend is toward leveraging the processing power of data platforms instead of siloed analytics tools.

A MODERN CLOUD DATA WAREHOUSE ALSO SUPPORTS MULTISTRUCTURED DATA. Structured data continues to be relevant, but data scientists and analysts are increasingly tapping multistructured data for ML and other analytics, especially in the form of JSON, XML, AVRO, PARQ, ORC, and a variety of log formats.

CONSIDER A MODERN CLOUD DATA WAREHOUSE AS YOUR PRIMARY PLATFORM FOR ML DATA CONTINUED

A CLOUD DATA WAREHOUSE PROVIDES A HOME FOR ALL ML DATA. Given the diversity of data seen across ML life cycle stages, data scientists run the risks of missing siloed data sets and losing control from a governance and curation viewpoint. Plus, ML's relentless data wrangling burns up expensive data scientist payroll and distracts them from their mandate of creating the best analytics possible. Organizations can avoid these problems by consolidating all data for ML, AI, predictive, and related practices into a modern, built-for-the-cloud data warehouse platform.



3

CONSIDER A DATA LAKE AS A SECONDARY PLATFORM THAT SATISFIES MULTIPLE ML DATA REQUIREMENTS

This report has already discussed the importance of learning and training data—terabyte-scale data sets of diverse data consumed during the design phase of a machine learning solution. There are many ways to provision data for machine learning. However, the trend is toward consolidating as much data as possible into a data lake, which integrates with and extends a data warehouse. Furthermore, data lakes are trending toward elastic clouds for their speed, scale, and economics.

In another related trend, data warehousing and analytics are trending toward multiplatform data architectures, which include a mix of big data platforms, clouds, data lakes, and data warehouses. This portfolio of diverse data engines and tools is increasingly *hybrid* in the sense that some systems and data are in the cloud and others are on premises.

Multiple types of data platforms integrate and work together in today's hybrid, multiplatform,

and multicloud data architectures. The advantage for users is that they can pick and choose within a diverse platform portfolio to get the most appropriate storage, read characteristics, in-place processing, and economics for each data set and analytics use case, including features optimized for ML and predictive analytics.

For example, the 2019 TDWI survey about cloud data management reveals that most of the tools and platforms used for data warehousing and analytics are already deployed in hybrid architectures. The bars on the left side of Figure 2 state the obvious; on-premises systems are still prominent. However, the bars on the right show that systems are also well-established on cloud. In addition, the bars overlap in the middle of Figure 2, indicating that warehouse and analytics platforms and tools are sometimes deployed as hybrid architectures.

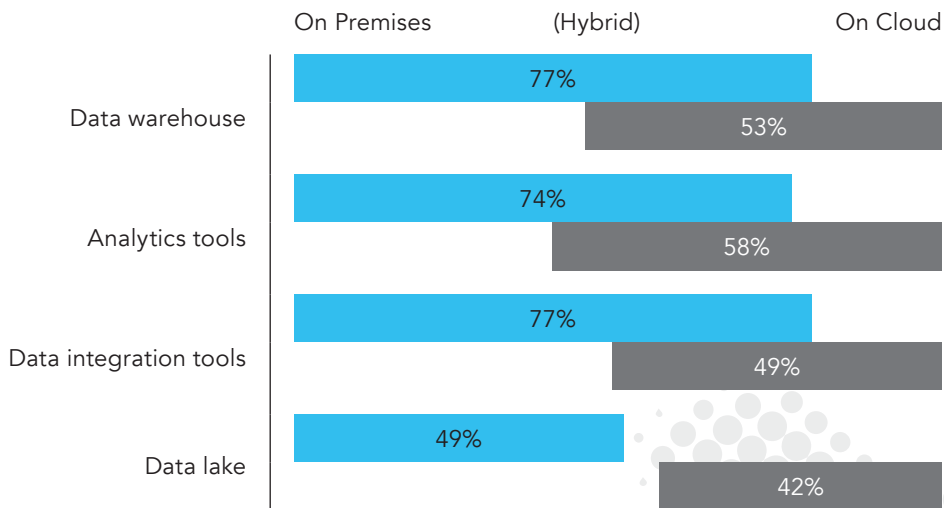


FIGURE 2. Systems architectures today are often hybrid, spanning on premises and cloud.³

³ Source: 2019 TDWI Best Practices Report: Cloud Data Management, Figure 16. Online at tdwi.org/bpreports.

CONSIDER A DATA LAKE AS A SECONDARY PLATFORM THAT SATISFIES MULTIPLE ML DATA REQUIREMENTS CONTINUED

A DATA LAKE CAN ENABLE A NUMBER OF PRACTICES ASSOCIATED WITH ML, AI, AND PREDICTIVE ANALYTICS:

- The data lake is defined as a repository for raw, detailed source data. Preserving data in its original state enables users to repurpose data repeatedly for a wide range of analytics and other use cases. This focus on raw data is also a good fit for ML's exploratory, learning, and training data during development, as well as capturing ML's production data.
 - Most data lakes are designed to enable self-service data exploration, prep, and visualization by data scientists and other technical users, as well as some mildly technical business users. However, this design frequently requires ETL or ELT processes on top of the data lake before data exploration and discovery can be performed during the first life cycle stage of ML.
 - Data scientists tend to generate many analytics sandboxes as they work, which is why well-governed data lakes have tools and data curation policies for sandboxes. This also keeps a lake from deteriorating into a data swamp. Sandboxing has direct application to ML's exploratory, learning, and training data.
 - A data lake platform can be tuned to capture the output of ML, including metadata on model drift in later life cycle stages, to monitor for drift and to retrain ML-driven predictive models.
 - A data lake can be deployed on cloud platforms, as with the modern data warehouse, for scale and speed with minimal cost.
- Data lake users think of their lake as part of their multiplatform data warehouse architecture, or they think of the two as separate but unified via many integration processes. Either way, between the lake and the warehouse, all data requirements of ML, AI, and predictive analytics can be satisfied optimally.



4

NOTE THAT DATA INTEGRATION IS A CRITICAL SUCCESS FACTOR FOR ML**DATA INTEGRATION PUTS THE RIGHT DATA IN THE RIGHT CONDITION ON THE BEST DATA PLATFORM ON A PER-USE-CASE BASIS.**

We have seen that each stage of the machine learning life cycle has distinct data requirements. We have also seen the important role played by data warehouses, data lakes, cloud, Hadoop, and object storage. However, for that much data to get that far after that much transformation requires substantial infrastructure for data integration (DI) and data quality (DQ), plus related disciplines in ETL/ELT, data virtualization, event processing, metadata management, and streaming ingestion.

After all, successful ML, AI, and predictive analytics need data that is prepared and presented in a certain way so that an analytics tool has data in a schema it can read optimally, in a quality condition that will not skew model designs or analytics outcomes, and on a data platform that has the functionality and performance characteristics that a specific analytics tool or user practice demands.

THE BROADER THE DATA, THE MORE COMPREHENSIVE THE ML-DRIVEN MODEL IS.

In other words, a data integration infrastructure should provision large volumes of diverse data for exploration, learning, and training stages—integrated from multiple, diverse sources with diverse latencies and representing various business entities, processes, and timescales. The breadth of data makes analytics or predictive models and assessments more real-world, accurate, and successful in production. Hence, investments in data integration and related disciplines are worthwhile because they raise the effectiveness of the resulting predictive model, which in turn raises the ROI of programs for big data, analytics, data warehousing, data lakes, and the automation of business operations.

DATA INTEGRATION SUPPORTS SPECIAL FUNCTIONS AND CHARACTERISTICS THAT CONTRIBUTE TO SUCCESSFUL ML, AI, AND PREDICTIVE ANALYTICS:

- **INTERFACES TO ALL SYSTEMS, BOTH SOURCES AND TARGETS.** To reach today's diversity of sources and the many targets seen in multiplatform data architectures, data integration tooling needs a large and mature library of interfaces. Luckily, DI and DQ tools have kept pace with the growing number of new source types (including SaaS-based operational applications and machinery or sensors in the Internet of Things (IoT)) and new target types (Hadoop or cloud databases, warehouses, and storage).
- **PUSH-DOWN AND IN-PLACE DATA PROCESSING.** Today's data integration and quality tool suites are feature-rich and powerful, yet they still benefit from leveraging the processing engines built into modern databases, cloud data warehouse platforms, Hadoop, and other data platforms. This takes many forms, including ELT push-down, in-place processing, and in-database analytics in multiple source languages (SQL, R, Python, etc.). This and other in-database processing helps data integration and quality to scale and perform, just as they provide similar speed and scale for model creation, updates, and scoring. Furthermore, processing data in place reduces the amount of data moved and simplifies data integration solution design.

**NOTE THAT DATA INTEGRATION IS A CRITICAL SUCCESS FACTOR FOR ML
CONTINUED**

- **DATA QUALITY FOR LEARNING AND TRAINING DATA.** As with all data-driven solutions, ML, AI, and predictive analytics are subject to the old adage: garbage in, garbage out. Hence, ML's learning and training data need data quality functions.

For example, tools capable of AutoML are regularly optimized for a specific schema; a DQ data standardization function can prepare training data appropriately. As other examples, outliers and redundancy in learning or training data can skew predictive model design and performance. A DQ deduplication function can find data copies and merge them, whereas a DQ profiling function can find outliers and normalize them.

- **REAL-TIME DATA CAPTURE FOR REAL-WORLD PROCESSES:** Some machine learning solutions compare the latest data to a predictive model as the data streams in real time. This requires that the data integration infrastructure include special functions or additional tools for handling data streams, event processing, and IoT.

- **DATA INTEGRATION CAN UNIFY THE MULTIPLATFORM DATA ARCHITECTURES TYPICAL OF MODERN DATA WAREHOUSING AND DATA LAKES.** Obviously, getting source data into complex architectures consisting of many data targets requires data integration. Not so obvious is the fact that data moves relentlessly among target platforms in a multiplatform architecture. This is to assemble sandboxes and other data sets as new analytics solutions are created, as well as to repurpose data for different use cases in analytics, reporting, business monitoring, and self-service practices. Similarly, machine learning data is moved and transformed as it moves through life cycle stages for exploration, learning, training, production, and model updates. Data integration assists with all these data movement and transformation scenarios—and more.



5

REMEMBER THAT ANALYTICS TOOLS HAVE DATA REQUIREMENTS, TOO

ANALYTICS BASED ON SETS VERSUS ALGORITHMS.

TDWI distinguishes between the general approaches of “set-based analytics” (e.g., OLAP, ad hoc queries, and 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. The broad and open-ended discovery and model creation tasks that we assume of ML development can work with either approach. However, the point is that analytics tools tend to support one or the other approach, and you will need to select tools and/or prepare analytics data accordingly.

IN-DATABASE ANALYTICS. The modern way to handle analytics data is to process it in place, in the database or other platform where it is stored. This alleviates the need to move massive amounts of big data to the analytics tool; instead, nowadays a tool’s algorithms and/or set operations are executed by engines built into the data platform. That, in turn, influences the data platform(s) that you would match to the requirements for ML and other analytics. In other words, look for data platforms that can execute the types of in-database analytics processing that you anticipate using. Likewise, make sure your data platform plays well with your chosen ML tools and vendors.

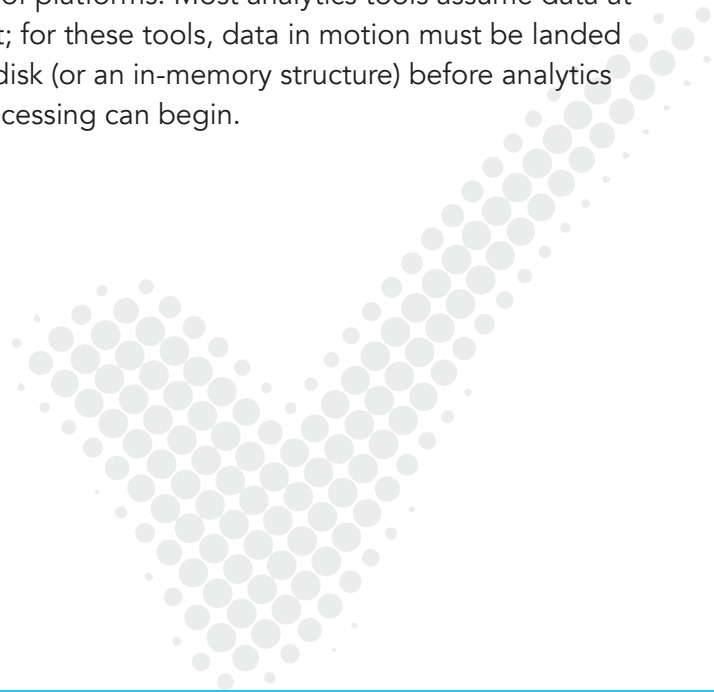
HAND-CODED PROGRAMS OR CODE GENERATION IN SPECIFIC LANGUAGES.

Development environments and their resources vary considerably. Some analytics tools are essential studios for hand coding in specific languages (commonly Java, R, and Python) as well as SQL routines pushed down into database management systems. These assume deployment to environments that can compile

or interpret code written in these and similar languages. This will affect your choice and data platform and how in-database processing is done.

STRUCTURED QUERY LANGUAGE. SQL is still very much relevant for analytics processing, particularly with relational data and set operations. Some vendors even support using SQL to natively query semistructured JSON, XML, PARQ, and AVRO data. SQL for ML lets you leverage the speed, scale, and flexibility of new cloud-based relational databases, as well as leverage existing team skills for creating solutions and optimizing them based on a familiar data language. In some cases (set-based instead of algorithmic ones), SQL processing is faster, more scalable, and simpler (less than 100 lines of SQL code versus thousands in Scala).

MISCELLANEOUS TOOL REQUIREMENTS. Some analytics tools require file-based data. A tool may be optimized for specific file and document types, such as web logs, XML, JSON, and Hadoop files. Other tools demand a flat file, perhaps with a recurring record structure. Some do not support any form of in-database analytics or they support a short list of platforms. Most analytics tools assume data at rest; for these tools, data in motion must be landed to disk (or an in-memory structure) before analytics processing can begin.



6

SELECT YOUR FIRST STEPS IN ML-DRIVEN AI AND PREDICTIVE ANALYTICS BASED ON BUSINESS NEED AND SENSIBLE PROJECT PLANNING

A 2019 TDWI survey about AI and ML asked: “What do you believe are the best practices to get AI projects off the ground?” The responses provide solid advice for taking the first steps in ML-driven AI and other predictive analytics projects (see Figure 3).

SELECT A CAUSE TO CHAMPION VIA ML AND AI.

Being a rebel without a cause rarely works in business, whereas championing a cause works well. In that spirit, over half of survey respondents think it best to “start with a business problem” (53%), such as fraud or customer churn. ML is the perfect solution to predict these events before they happen while there is still time to stop them and establish immediate business value for the investment in analytics. A business opportunity would make a similar successful start, such as automating the decision-making components in sales or customer service processes. These are examples of “achievable projects where [ML-driven] AI can help.”

MARSHALL ORGANIZATIONAL RESOURCES.

“Executive sponsorship” (40%) can take various forms, including a business manager willing to spend resources on new analytics or a chief officer with a big stick who feels that ML/AI and

other analytics are critical to the firm’s future. To assure that the resulting analytics solution aligns to business goals, “make sure business and technology groups collaborate” (34%).

PLAN THE ML/AI PROJECT CAREFULLY. A best practice in project management is to “deploy an important, high-impact project [phase] first” (17%). This way, an early success builds confidence among business and technology people alike, without which the project may not be allowed to continue. Similarly, it is best to “deploy AI technology in an already existing process” (16%). Creating a new business process and its embedded analytics early in the program is too much of a big bang project, which would be prone to failure due to rolling out too much functionality too early.

MAKE SURE DATA IS HIGH QUALITY (33%). We’ve already noted how poor data quality can skew the design and performance of ML-driven predictive models. In addition, discovering and correcting data quality can easily become its own large project, thereby stealing time and resources from the analytics solution. If addressing data quality is inevitable, then at least see it coming and plan accordingly by deeply profiling important data sources and prototype data sets for ML learning and

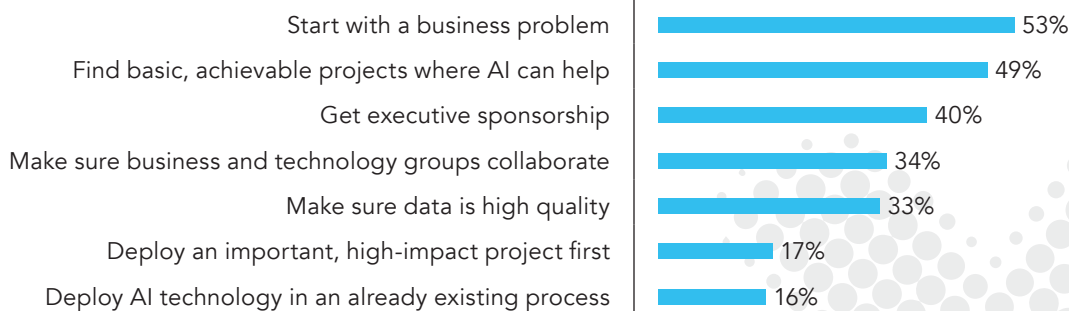


FIGURE 3. Based on 256 respondents.⁴

⁴ Source: 2019 TDWI Best Practices Report: Driving Digital Transformation Using AI and Machine Learning, Figure 11. Online at tdwi.org/bpreports.

**SELECT YOUR FIRST STEPS IN ML-DRIVEN AI AND PREDICTIVE ANALYTICS
BASED ON BUSINESS NEED AND SENSIBLE PROJECT PLANNING CONTINUED**

training. In a similar vein, ensure that the rest of the required data management infrastructure is in place before ML-driven development begins, namely data integration and quality tooling and data platforms that will persist data for machine learning. After all, this comprehensive data management infrastructure is required to satisfy all the data requirements of ML's life cycle stages, namely the creation and management of exploratory data, learning data, training data, production data, and retraining data.



ABOUT OUR SPONSOR



Snowflake started with a clear vision: make modern data warehousing effective, affordable, and accessible to all data users. Snowflake enables the data-driven enterprise with instant elasticity, secure data sharing, and per-second pricing, across multiple clouds. Because traditional on-premises and cloud solutions struggle at this, Snowflake developed a new product with a new built-for-the-cloud architecture that combines the power of data warehousing, the flexibility of big data platforms, and the elasticity of the cloud at a fraction of the cost of traditional solutions. Snowflake: your data, no limits. Find out more at snowflake.com.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

ABOUT 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, having published over 600 research reports, magazine articles, opinion columns, and speeches over a 20-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 was a product manager at database vendors. His Ph.D. is from Yale. You can reach him by email (prussom@tdwi.org), on Twitter ([@prussom](https://twitter.com/prussom)), and on LinkedIn (linkedin.com/in/prussom)

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

