research

CHECKLIST REPORT

2018

The Automation and Optimization of Advanced Analytics Based on Machine Learning

By Philip Russom

Sponsored by:

Qubole®

tdwi

Transforming Data With Intelligence™

MAY 2018

TDWI CHECKLIST REPORT

The Automation and Optimization of Advanced Analytics Based on Machine Learning

By Philip Russom

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

TABLE OF CONTENTS

FOREWORD

Several trends are driving organizations toward machine learning.

As with other forms of advanced analytics, the basic concepts and techniques of machine learning have been around for decades. However, a number of trends have converged to make machine learning (ML) suddenly more desirable and practical than ever before:

- **User organizations need a wider range of analytics.** The drive to be more competitive, profitable, agile, innovative, and growth oriented has spurred organizations to investigate new approaches, including predictive analytics (as enabled by machine learning).
- **Moore's Law has taken us to a higher level of speed and scale.** These improvements are required to get value from big data and other voluminous data sources, such as social media and the Internet of Things (IoT). High performance and ample training data accelerate the development of analytics models and algorithms, as well as the continuous tuning of these via machine learning during deployment.
- **Analytics tools are better than ever.** The vendor and open source communities have given us greater ease of use and design depth for the automation and optimization of machine learning.
- **Data professionals are addressing the skills gap.** Driven by business demand, technical staff and some data-savvy business users have developed new skills for data science, big data, and advanced forms of analytics such as machine learning. Closing the skills gap is a critical success factor for organizations seeking to compete on analytics and to leverage big data for organizational advantage.

Machine learning algorithms learn from large data sets to create predictive models.

Here's how it works: Machine learning algorithms consume and process large volumes of data to learn complex patterns about people, business processes, transactions, events, and so on. This intelligence is then incorporated into a predictive model. Comparisons to the model can reveal whether an entity is operating within acceptable parameters or is an anomaly. Machine learning is being used today to solve well-bounded tasks such as classification and clustering. Note that a machine learning algorithm learns from so-called training data during development; it also learns continuously from real-world data during deployment so the algorithm can improve its model with experience.[1]

Machine learning has serious data requirements that are critical to success.

- **Machine learning demands large, diverse data sets.** Prior to model design, a machine learning algorithm's learning process depends on large volumes of data, from which it draws many entities, relationships, and clusters. Volume aside, data integrated from diverse sources tends to broaden and enrich the correlations made by the algorithm.
- **Large data sets demand large, diverse infrastructure for data management.** Infrastructure for machine learning's training data typically involves multiple data platforms, tools, and processing engines, ranging from traditional (relational and columnar databases) to modern (Hadoop, Spark, and cloud storage). Multiple technologies are required to cope with training data's extreme size, multiple data structures, and (in some cases) multiple latencies. In short, tools for machine learning are obviously important, but data management infrastructure is just as important.

This report will drill into the data, tool, and platform requirements for machine learning with a focus on automating and optimizing ML's development environment, production systems, voracious appetite for data, and actionable output. The point is to provide useful information for organizations that need machine learning for business analytics and also need to get greater business value from big data and other new data sources.

[1] This definition of ML paraphrases two TDWI publications. For more information, see the 2017 *TDWI Best Practices Report: Advanced Analytics: Moving Toward AI, Machine Learning, and Natural Language Processing* and the 2018 *TDWI Checklist Report: Seven Best Practices for Machine Learning on a Data Lake*. Both are available at tdwi.org.

NUMBER ONE

GIVE MACHINE LEARNING A DATA ENVIRONMENT THAT IS AS DIVERSE AS IT IS BIG

Machine learning has a voracious appetite for data during both development and production, which makes unique demands of an organization's data management infrastructure.

- **Machine learning is small but powerful.** ML algorithms and models are tiny compared to the vastness of big data and the multiplatform infrastructure required to capture and leverage it. The broader the data (in terms of its sources and the entities represented), the more comprehensive the model is. Hence investments in data management are worthwhile because ML provides insights that raise the ROI of programs for big data, analytics, data warehousing, data lakes, and so on. Furthermore, when organizations already have a big data infrastructure, adding ML extends the life cycle and business value of that infrastructure.
- **Data management infrastructure can be vast.** It can, for example, include platforms and tools for data warehousing, data lakes, data integration, data preparation, multiple forms of analytics, and big data. New data platforms are emerging as well, dominated by clouds; open source engines, libraries, and languages (for example, Apache Spark); and self-service tools. That is a long list of platforms, technologies, and processing engines. Yet, it is all required for modern organizations that want to operate and compete on analytics and intelligence.
- **Each form of analytics (including ML) has its own data requirements.** First, savvy organizations are deploying tools for multiple types of analytics (not just machine learning) because each type tells them something unique and valuable. Second, each analytics approach needs data that is prepared and presented in a certain way so that an analytics tool has data in a schema, a quality condition, and on a data platform optimal for the analytics tool or the user practice involved. For example, machine learning algorithms are almost always optimized for raw detailed source data. Thus, the data environment must provision large quantities of raw data because that is required for discovery-oriented analytics practices such as data exploration, data mining, statistics, and machine learning.
- **ML needs data from diverse sources, in diverse formats, about diverse business processes.** For the most comprehensive learning experience, a data management infrastructure should provide diverse training data—integrated from multiple, diverse sources and concerning various business entities—to make algorithmic assessments more real-world, accurate, and successful in production.

To support machine learning successfully, the data management infrastructure needs speed and scale.

- **Scale to large volumes of training and test data:** A common reason for model failure in production is the lack of proper training data, which needs to be massive, diverse, and from real-world processes.
- **Speed and scale for analytics processing:** Machine learning is very good at finding patterns in large amounts of data. For instance, models built with machine learning can establish baseline profiles for various entities (e.g., authorized versus unauthorized transactions) and then predict which users or accounts are likely to present unauthorized transactions. This approach must scale and perform with both baseline creation and production comparisons, even in multiterabyte data environments.
- **Speed for agile, iterative development:** Machine learning design usually requires an iterative process, where a developer tweaks an algorithm and reruns it immediately. The data environment must perform with low latency, even with large data sets and distributed queries that hit multiple systems.
- **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 management infrastructure include special tools and platforms for streams or event processing.

NUMBER TWO

EMBRACE DATA TECHNOLOGIES AND APPROACHES THAT ARE KEY TO MACHINE LEARNING SUCCESS, NAMELY DATA LAKES AND CLOUDS

Multitechnology infrastructure has become the norm for data environments, especially in analytics-driven organizations.

This is because it takes diverse platforms, processing engines, tools, and technologies to satisfy the requirements of diverse data and diverse analytics. The result is today's multiplatform data management infrastructure, which—in modern firms—includes 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 on clouds and others are on-premises. All these data platforms work together in today's hybrid, multitechnology data environments and users can

pick and choose among them to get the most appropriate storage, read characteristics, in-place processing, and economics for each analytics use case, including ML for predictive analytics.

Finally, note that this large data management infrastructure does not exist for ML alone; it also supports other forms of analytics, reporting, data preparation, data exploration, self-service data practices, and even operational applications.

Analytics users are trending toward data lakes and clouds.

This report has already discussed the importance of 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 training data for machine learning. This data can come from multiple platforms in the extended data infrastructure—typically enterprise applications, data warehouses, IoT sources, clouds, and lakes. However, the trend is toward consolidating as much data as possible into a data lake. Furthermore, data lakes are trending toward elastic clouds for reasons of automation, optimization, and economics.

- **The data lake manages the massive volume of detailed source data that ML needs.** In fact, that's what a data lake is—a large repository of raw data of which most is captured and persisted in its original state as it comes from a source system or stream. Such a raw data repository is nirvana for analytics users. They can return to the original data over and over to repurpose it as new business questions or analytics projects come along. This isn't possible with data warehouses, which contain mostly aggregated and calculated values for reporting.

  Another benefit of the data lake is that there's no need to move data. The data needed for ML is already in the lake and data sets generated by ML processing can be stored there as well. For these reasons, the data lake—though only a few years old—has quickly become the preferred big data store for discovery-oriented analytics ranging from self-service data preparation and visualization to data mining and machine learning.

- **The cloud is ideal for the demanding and unpredictable workloads of ML and other analytics.** For example, the iterative reads of large data sets that are common with ML development would wreak havoc with traditional data warehouses and other relational environments. Clouds cope with these easily using elasticity, in the form of workload-aware auto-scaling, which marshals resources as workloads ramp up, then reallocates resources as loads subside.

  As another example, building a large data lake with traditional platforms is cost prohibitive, but a data lake on modern cloud storage is comparatively cheap. Furthermore, an on-premises data lake demands time-consuming and risky system integration, whereas cloud providers handle the system integration for you. For these reasons, the cloud has become the preferred medium for data lakes and all forms of analytics, including machine learning.

NUMBER THREE

CONSIDER MAKING APACHE SPARK YOUR PREFERRED ENGINE FOR MACHINE LEARNING

Apache Spark is an open source clustered computing framework for fast and flexible large-scale data analytics. TDWI sees Apache Spark as an important new technology, especially for analytics, big data environments, and Web-based operations.

Spark offers many benefits for analytics and machine learning:

Spark integrates with many data platforms and related systems. These include platforms important to ML and other analytics such as Hadoop, cloud-based object storage, and popular cloud-based data warehouse platforms.

Spark integrates with the rich ecosystem of Apache tools. Users can take advantage of multiple engines, tools, and languages, then select the best tool for a given use case, data set, or workload.

Spark primitives operate in-memory. For example, Spark has a parallel data processing framework that places data in Resilient Distributed Data Sets (RDDs), a distributed data abstraction that scales to complex calculations with fault tolerance. This elimination of input and output (I/O) provides speed for iterative analytics (as with ML development) and modern data pipelining. Furthermore, once data is in Spark memory, many tools and applications can access it easily at high performance.

Spark's architecture decouples compute and storage resources. This contributes to greater speed and scale for Spark clusters, especially compared to Hadoop clusters, which are not decoupled. Decoupling also gives developers greater flexibility with deployment designs. Furthermore, Spark clusters can be heterogeneous as well as homogeneous.

Spark libraries are highly useful to data management and analytics specialists. This is especially true of Spark's libraries for standard SQL, GraphX, and machine learning (called MLLib).

ML users should look for tools with deep support of Spark.

Look for tools that can automatically spawn Spark clusters. This automation simplifies and accelerates incorporating new sources and the use of their data assets.

A tool should help Spark auto-scale based on workload recognition. This optimization simplifies resource allocation and management so scalability is reached sooner and more easily. It also assures ample resources for analytics workloads such as ML.

A tool should provide extra security for the Spark cluster. For example, look for encrypted credentials.

Many Spark users prefer open source notebooks. Look for tools that are compatible with Zeppelin and Jupyter.

NUMBER FOUR

ACHIEVE SPEED AND SCALE—KEY REQUIREMENTS FOR ML DESIGN AND OPERATION—BY ADOPTING NEW ARCHITECTURES

Hadoop has served a useful purpose by being a big data platform that users could afford and learn on. However, dissatisfaction is mounting because of Hadoop's limitations in speed, security, metadata, and SQL compatibility. Furthermore, Hadoop architecture tightly couples compute and storage resources, which means you cannot scale up resources in one area without also doing so in the other.

Compute and Storage Decoupled

For better resource management and therefore better scaling, the current trend is toward big data platforms that decouple compute and storage. The trend is especially apparent in the growing adoption of other open source engines, such as Apache Spark and cloud-based object storage.

When compute and storage are decoupled, they can be managed as separate resources. This allows compute and storage to scale and perform independently instead of being handcuffed together. In turn, this independence reduces limitations, resulting in greater speed and scale.

Decoupling has positive ramifications for ML. The performance and scale improvements of decoupling empower ML algorithms and models to read and score larger data volumes within smaller timeframes. In other words, data teams can run more jobs with the same budget, thereby keeping total cost of ownership low. Furthermore, decoupling provides more flexibility in how system resources are managed, so that developers can be more innovative in designing solutions for analytics and ML.

Apache Spark Architecture

Spark's data platform architecture has scalability, low cost, and compatibility with the Apache tool ecosystem similar to Hadoop but without Hadoop's limitations. For example, Spark has the linear scalability of MapReduce but with high performance and low latency. Spark enables iterative development (an ML requirement) and ad hoc queries with big data that Hadoop users can only dream of. Spark's library architecture means it will grow into ever-broader functionality, including a library for ML. Spark can integrate with Hadoop to improve it, but Spark can also work with many other data platforms, including cloud storage.

Cloud-Based Object Storage

Cloud storage has matured recently to support strategies based on objects, blocks, or files and folders. Among these, object storage is preferred for content-driven interfacing, as is typical with machine learning and some other analytics methods. The value of object storage lies in simplicity (it's easy to work with), economy (object storage tends to be cheaper than block storage), and its not being coupled with compute (thereby removing limits to speed and scale).

Spark and Object Store as an Integrated Architecture

Data and analytics professionals have started integrating Spark and object storage. TDWI expects this to become a common architecture because both Spark and object storage have compelling functionality for analytics and data management (as discussed throughout this report) and both decouple compute and storage for maximum speed, scale, and flexibility.

NUMBER FIVE

EMBRACE DATAOPS AND OTHER MODERN TEAM STRUCTURES THAT AFFECT MACHINE LEARNING SUCCESS

Until recently, data-driven development was slow, siloed, and misaligned.

To get an ML model from inception to production, someone must collect data of interest for the project, explore data looking for insight, firm up a hypothesis based on what they discovered, collect more data based on the hypothesis, create an initial prototype of a model, get feedback about the prototype, iteratively evolve the model, collect even more data, get more feedback, iterate the prototype further, try out the model in a test system, revise the model accordingly, deploy the model in a production system, test again, revise again, and finally release the model (or its parent solution) to users.

Whew! That's a lot of steps and all of them are complex and time-consuming. Although some steps are performed by the same person, there are still many people required to complete the process. This situation—until recently—was exacerbated by the fact that technical personnel for data, analytics, testing, and production were on different teams, each fairly siloed. Communications among teams was slow, inaccurate, alienated from business goals, and a barrier to innovation.

DataOps cures team silos for better collaboration, speed, and alignment.

Outside of data management and analytics, applications developers had similar silo problems. They cured them with DevOps, a practice of combining software engineering, quality assurance (QA), and operations into a single, agile, and collaborative team structure. Very recently, data and analytics people have adapted the principles of DevOps to create DataOps, a new way of managing data that promotes communication between—and integration of—formerly siloed data, teams, tools, technologies, and platforms. DataOps fosters collaboration among everyone handling data, including data developers, engineers, and scientists, as well as analysts and businesspeople.

Let's take another look at the ML model development process as described at the beginning of this section to see how DataOps simplifies, speeds up, and aligns ML development and production:

- Days and weeks of downtime may pass between adjacent steps. The collaborative relationships created by DataOps among technical team members reduce the downtime so that ML models and other solutions get into users' hands much sooner.
- In many cases, independent teams work from specifications that are impossible to keep current. The documentation approach—created for data modeling—does not adapt well to analytics modeling. Luckily, the direct communication that DataOps fosters eliminates the need for time-consuming, misleading, and distracting specs and documentation.
- Ideally, the unified DataOps team also has collaborative relationships with businesspeople who are committed to fast and frequent reviews of prototypes and iterations. With the right businesspeople involved at many stages in the process, building an analytics model that aligns with business goals is far more likely.
- An extended DataOps team concentrates substantial domain expertise—from analytics to enterprise data management—that can quickly and easily be tapped for the accelerated development of data-driven products. For example, with machine learning, a data scientist may lead development but be supported by colleagues who become the stewards of the data pipes that data scientists need for their ML algorithms.

Other modern team structures can accelerate machine learning development.

Note that there are other modern team structures and methods that achieve results similar to those of DataOps in improving the speed, efficiency, standards, and alignment of data-driven development and related services:

- **The Agile Manifesto.** This was originally written as a method for application developers. Its adaptation to data practices has revolutionized analytics and data set development, making them leaner, nimbler, and better aligned with business goals.
- **Data stewardship.** Originally a way for businesspeople to collaboratively guide data quality projects, stewardship has been adapted to data warehousing, data integration, and master data management projects. To assure analytics-to-business alignment, a data steward can provide business requirements and review iterative versions of analytics models.
- **Data competency centers (or centers of excellence).** These consolidate siloed data-driven teams into a single, centrally managed team based on a shared-services model. Specifics vary greatly, but most competency centers also enforce enterprise standards for data and are strongly allied with data governance efforts.

NUMBER SIX

REMEMBER THAT MACHINE LEARNING IS NOT JUST FOR PREDICTIVE ANALYTICS

Today, most efforts with machine learning are to support predictive analytics, especially when the analytics parses vast amounts of diverse big data. This is an important practice and it will continue to grow and mature.

However, a few cutting-edge vendors and open source projects are embedding ML-driven intelligence into data management (DM) tools. Embedded within these DM tools, ML algorithms and models typically address three broad goals:

- **Automation** for well-understood but time-consuming development tasks such as mapping sources to targets, cataloging data, or onboarding new sources.
- **Optimization** of system performance by automatically selecting query optimization strategies, table join approaches, resource management schemes, and distribution methods for data (e.g., hot versus cold storage, memory versus disk, or replication across nodes).

- **Capacity management** via workload-aware auto-scaling, spot instance purchasing, and integrating node types in heterogeneous clusters.

Machine learning is high value in these contexts because it increases developer productivity, makes advanced functions doable by lightly technical users, and elevates system performance with minimal administrator involvement. Due to these compelling benefits, TDWI expects to see most DM functions automated or optimized via ML and other approaches (e.g., rules engines) within a few years. Here are a few examples:

Data cataloging. Modern tools can catalog and categorize data automatically via ML algorithms and models, as well as via old-school business rules and application logic. Cataloging can apply to data sources, data sets, tables, or even individual columns and fields. A single data element can be categorized by its domain, compliance risk, quality level, source, lineage, and so on, as the user organization requires. Cataloging each data element multiple ways enriches user searches and queries of the catalog, as well as enabling richer cross-category correlations.

Data domains. ML algorithms and other tool logic can recognize and catalog data sources and structures of particular domains. This helps users browse or search the catalog for domains of high interest, such as the customer, product, and financial domains. Advanced algorithms can even detect domains and domain relationships across data sets. ML algorithms can also recognize and catalog data elements that are potentially sensitive in terms of privacy and compliance.

Data lineage. ML algorithms can parse large volumes of complex data (even data distributed across multiple data platforms) to record data pathways and cluster data elements and data sets of common origin. With these details, users can quickly obtain deep insights into data provenance and impact analysis.

Metadata management. With big data, IoT, and other new sources that are notoriously devoid of metadata, a modern DM tool with ML embedded can parse data and deduce credible metadata. The tool can then suggest a metadata structure to a data developer for approval or log that structure in a metadata repository without human intervention.

Data mappings. Time-consuming source-to-target mappings can now be performed by ML models and algorithms. ML's accuracy and breadth increase as it watches users manually map successfully. Automated mappings increase the productivity of data developers, data scientists, and data-savvy business users.

Data anomaly detection. ML has the potential to spot and react to data defects such outliers, nonstandard data, and various data quality issues. Some tools go beyond detection to automatically remediate data quality issues based on ML models or encoded business rules.

Upcoming use cases for the ML automation and optimization of DM. In the near future, catalog-based ML will also contribute to data security, governance, capacity planning, system performance, and guided data exploration.

ABOUT OUR SPONSOR

qubole.com

Qubole is revolutionizing the way companies activate their data—the process of putting data into active use across their organizations. With Qubole's cloud-native Big Data Activation Platform, companies exponentially activate petabytes of data faster, for everyone and any use case, while continuously lowering costs. Qubole overcomes the challenges of expanding users, use cases, and variety and volume of data while constrained by limited budgets and a global shortage of big data skills. Qubole's intelligent automation and self-service supercharge productivity, while workload-aware auto-scaling and real-time spot buying drive down compute costs dramatically. Qubole offers the only platform that delivers freedom of choice, eliminating legacy lock in—use any engine, any tool, and any cloud to match your company's needs.

Qubole investors include CRV, Harmony Partners, IVP, Lightspeed Venture Partners, Norwest Venture Partners, and Singtel Innov8. For more information visit www.qubole.com.

ABOUT THE AUTHOR

Philip Russom, Ph.D., is TDWI's senior research director for data management and oversees many of TDWI's research-oriented publications, services, and events. He is a well-known figure in data warehousing and business intelligence, having published over 500 research reports, magazine articles, opinion columns, speeches, Webinars, and more. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and BI consultant and was a contributing editor with leading IT magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on 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 TDWI CHECKLIST REPORTS

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