research

CHECKLIST REPORT 2020

Evolving Data Science in the Cloud

Five factors for success

By Fern Halper, Ph.D.

Sponsored by:

sas

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

NOVEMBER 2020

TDWI CHECKLIST REPORT

Evolving Data Science in the Cloud

Five factors for success

By Fern Halper, Ph.D.

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

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

Analytics is evolving. For example, at TDWI we see that self-service is a top priority among organizations. We also see demand for more advanced analytics, such as machine learning and AI, is increasing. According to TDWI research, over 70 percent of survey respondents say demand for machine learning is growing in their organizations.[1]

Machine learning and other advanced analytics often utilize high volumes of diverse data for model training. To support modern analytics, organizations realize they must modernize their data infrastructure as well; supporting modern analytics is a top reason that organizations evolve a more advanced data management strategy. Typically, many models—hundreds or more—will be trained and evaluated in order to identify the one that works best for a particular problem. The traditional enterprise data warehouse was not designed to support high-volume, diverse data for computationally expensive iterative analytics, such as that required for machine learning.

Today, many organizations are creating multiplatform environments to support their analytics efforts. The cloud is an important piece of this strategy. In fact, TDWI research indicates that platforms such as cloud data warehouses or data lakes are a growth area for data management to support analytics. The cloud has numerous benefits for advanced analytics; top benefits include scalability and elasticity.

When you need to perform analytics processing on a large data set and iterate on that analysis, the cloud enables you to procure as much storage and compute services as necessary. When you are finished with the analysis, you are no longer responsible for those additional services. This is critical for analytics and especially for compute-intensive data science initiatives. For example, some organizations will invest in specialized compute resources to gain extraordinary performance improvements, such as the use of GPUs for deep learning. There is no reason to buy this hardware anymore; it's much easier to "borrow" it from a cloud provider.

Additionally, the end result of a data science effort is often a model that becomes part of an application that needs to run at scale. For instance, a fraud algorithm may be used for thousands or even hundreds of thousands of predictions per second. As organizations move to the cloud for analytics, they are examining how to design applications that are cloud-native and make use of the unique properties of the cloud. This often involves microservices and containers as part of the architectural plan.

This Checklist examines five best practices for utilizing the cloud for data science—including evaluating use cases to run in the cloud, cloud computing architectures, and planning considerations.

[1] Unpublished TDWI survey, 2020

1 EVALUATE YOUR ANALYTICS USE CASES

There is a tremendous amount of unrealized value in analytics use cases that never see the light of day. Organizations can no longer afford to leave this value on the table. The cloud makes it easier by removing computing as a barrier—an important consideration for many data science use cases.

If your organization is looking to utilize data science or is already using data science technologies, it should evaluate potential use cases where the cloud can help. These include:

- **COMPUTE-INTENSIVE MACHINE LEARNING TASKS.** Organizations are collecting increasing amounts of diverse data that is often generated in the cloud. These organizations are looking past building models that simply rely on structured data found in their on-premises warehouses. In fact, at TDWI we see that mainstream enterprises are already utilizing unstructured data such as text data. We see more organizations employing deep learning against large volumes of image data for applications from medicine to agriculture. Real-time applications are becoming more common. Training these models requires a large amount of compute. Organizations can use the scalable and elastic nature of the cloud: spin up the resources when needed, and spin down the resources once a model has been developed.
- **EDGE ANALYTICS.** Some use cases need to operate in the network, in real time. This is referred to as the edge: the point where traffic enters and leaves the network. For example, a utility company may want to monitor its wind turbines and determine whether a part is going to fail. It may make sense to capture the data close to the turbine and even analyze some of the data there. That can mean processing and analyzing the data at the edge rather than sending the data to a central system for analysis, which can be inefficient and risks bottlenecks. In this case, the analytics might happen in the device itself or in platforms near the edge in the cloud.
- **SANDBOX FOR EXPERIMENTATION.** Data science involves a lot of data exploration and experimentation. For instance, data scientists spend a lot of time on feature engineering—the process of building meaningful attributes from the data for model training. They may try different models and want to experiment. Often, organizations set up sandboxes for data scientists for this experimentation. It can make sense to put these sandboxes in the cloud, especially if the data is generated in the cloud (see the next page).
- **MODELS IN PRODUCTION.** Organizations want to deploy models into production. For instance, organizations may want to deploy a recommendation engine on a website or build an application that utilizes a machine learning model running on data created in the cloud. These applications may make sense to run in the cloud, although they may not run at the edge.

2 CONSIDER DATA GRAVITY

As mentioned, data science often uses complex data types such as images, machine data, or text data. For example, deep learning is now used to train models to identify weeds in a crop field or to determine if a crop needs watering. Drones collect this data. Sensors, RFID chips, scanners, and video produce other machine-generated data. Depending on the use case, this can involve massive amounts of data generated in real time. Much of it arrives in streams. Often, this data is generated in the cloud.

If the data is generated in the cloud, it makes sense to analyze it there too. This is the concept of data gravity; if the data is massive enough, it can be hard to move it.[2] In other words, it is important to leave the data where it is for analytics. If it is in a cloud data lake or platform, leave the data there and bring the compute to the data.

For data science, leaving data in place in the cloud for analysis can prove valuable for several reasons:

- **NO PERFORMANCE OVERHEAD.** If a data scientist is going to analyze large amounts of data that is generated in the cloud, it doesn't make sense to make a copy of the data and pull it down into some other storage area for analysis. Instead, it can be better to use the compute power of the cloud. This minimizes the latency issues of trying to move, for example, streaming data generated in the cloud to an on-premises location for further analysis. Instead, the organization keeps the data resources close to the processing resources.

- **NO EGRESS FEES.** Additionally, if the data is generated in the cloud and analyzed there, the data often doesn't need to leave the network topology. That means there are no egress fees for moving data out of the cloud. Data can be cheap to store in the cloud, but egress fees to move it out or sometimes between zones within the same cloud can be expensive. Once you land data in the cloud, in many instances, you don't want to move it around. The key is to move analytics compute to the data. The good news is analytics algorithms are often already part of most cloud environments, and analytics systems that are not part of the cloud can also be co-located with the data and deployed in the cloud alongside the data.

- **SECURITY CONCERNS.** Aside from cost and performance considerations, some data and analytics workloads need to be kept localized for security or compliance reasons. For instance, some data must remain within one local network, country, or region to meet compliance mandates. That data can be analyzed in place.

- **COMPUTE SERVICES.** It is often thought that if you keep data on your cloud tenant, you can't take advantage of compute services provided by others. However, the compute, regardless of the tenant, can sometimes be securely accessed as long as it is on the same cloud. In other words, if the data resides on the same cloud, organizations can take advantage of a variety of multitenancy configurations to ensure security while maintaining availability for more local analytics compute resources.

  For example, someone could have data in their Azure tenant, and allow applications managed by another vendor, such as an analytics vendor, on the same Azure cloud to work with that data. This means you don't have to use the cloud provider's exclusive services.

[2] The term was first coined by David McCrory in a 2010 blog post where he noted, "As Data accumulates (builds mass) there is a greater likelihood that additional Services and Applications will be attracted to this data." https://datagravitas.com/2010/12/07/data-gravity-in-the-clouds/

3 UNDERSTAND EVOLVING ARCHITECTURES FOR ANALYTICS

As an organization's data science efforts evolve, they will build more models and put them into production. This involves a software development mentality because a predictive model is really a piece of software and should be treated as such. Two important concepts for modern cloud architectures and data science are that cloud environments are often services-based and that they can utilize a container-based architecture. Microservices and containers (as well as Kubernetes) have redefined how model development and deployment can be accomplished.

MICROSERVICES. In a services world, applications are assembled using a set of loosely coupled services. The microservices approach segregates functionality into small autonomous services where each service has its own function. This is different from a traditional approach where the application has all of its functionality running in a single monolithic process. These services communicate via a well-defined interface using APIs. Microservices are easier to update, debug, and deploy.

Cloud service providers use services for numerous tasks such as networking, messaging, or logging. Microservices are important for data science as well, especially when it comes to deploying models. For instance, one service might compute a prediction. Another service might call the prediction. Microservices architectures make applications easier to scale and faster to develop. Each component of a machine learning-based application can be deployed and maintained as a self-contained microservice. If any element of that application needs to be updated or changed, only that element needs to be touched.

CONTAINERIZATION. Containerization involves packaging up software code and all its dependencies in a container so the software can run on practically any infrastructure. Containers are becoming popular. In a recent TDWI survey, for example, approximately 25 percent of respondents were already using containers.[3] (One popular container framework is Docker.)

Machine learning models, including the code, dependencies, tools, libraries, and configuration files, can all be packaged up into a container. Data scientists can then use containers to share their work and put it into production. This also means that others can rerun/reproduce the work. In the cloud, the container can be replicated to run across a cluster.

Microservices and containerization are the main tenets of designing for cloud-native compatibility. Kubernetes is the de facto standard for orchestrating these services. Kubernetes is an open source container orchestration platform developed by Google in 2014. The container-based architecture, orchestrated by Kubernetes, provides portability across different cloud environments, including Azure, Google, and AWS.

Additionally, because microservices/container-based architectures inherently need interoperability enabled by APIs, these kinds of cloud-native architectures are often more accessible from other applications. With APIs in place, it is easier to embed analytics services into operations.

[3] Unpublished TDWI survey, 2020

4 THINK ABOUT HOW TO OPERATIONALIZE DATA SCIENCE

When organizations operationalize analytics, they make it part of a business process, system, or application. Operationalizing analytics is key to success. Although organizations spend a lot of time thinking about building machine learning models, they often spend less time thinking about how they are going to deploy these models into production. It is important to think about model deployment early in the process.

TDWI research indicates the average number of models organizations have deployed in production is somewhere between two and five. Typically, organizations will manually manage these models. However, once organizations start to scale the number of models in production, they have to consider a number of factors, including how to manage the models in production.

Regardless of how the models are deployed (e.g., containers, APIs, etc.), it is important that models running in production be registered, tracked, and monitored.

Ideally, model management is centralized and there is full visibility and control over all modeling activities to ensure models are documented, the knowledge shared, and models can be tracked, monitored, and retrained.

Key functionality includes:

- **MODEL REGISTRATION AND DOCUMENTATION.** Just as you would version and document a piece of software, models need to be versioned and registered as well. This is vital during model training as well as when new versions of models are deployed after you've made a change.

  Registering models provides information (i.e., metadata) about who built it, when they built it, who has touched it, important attributes in the model, and other documentation. It also provides information about how many versions of the model have been built. Some organizations with only a few models use a manual process to register models. However, as the number of models created and changed grows, a manual process using a file directory will not be sustainable.

  Additionally, there is turnover in data science teams. If someone leaves the team, information about the model becomes that much more important. Otherwise, teams can spend a lot of time trying to figure out how the model works. Finally, some industries have audit and regulatory rules and guidelines about reproducibility, which also makes registering models key.

- **MODEL MONITORING.** Enterprises must monitor deployed models to see if they degrade (sometimes called drift) over time. Data changes. Business conditions change. Competitors change. Models must be updated to reflect this to continue performing optimally.

  For instance, a company might have created a supply chain model prior to the current COVID-19 crisis. The model predicts out-of-stock items in particular stores based on previous buy rates. However, if external conditions change, that model will not do a good job keeping shelves stocked.

- **MODEL RETRAINING.** As discussed, the past is not always going to be similar to what happens in the future. Data for models tends to drift from the original training set, often because the assumptions used to build the model have changed. An important piece of operationalizing analytics is retraining (and continuing to

THINK ABOUT HOW TO OPERATIONALIZE DATA SCIENCE CONTINUED

version and track) models once they've been in production and the organization is monitoring their performance.

Operationalizing advanced analytics is one reason why organizations utilize both commercial and open source solutions; open source solutions don't yet provide an effective way to deal with many of these issues. Sometimes a company might create a model using Python but then manage that model in production using a commercial product.

5 CREATE THE PLAN

For organizations to move forward utilizing the cloud for data science, they will need a plan. Typically, this is a phased plan that addresses the following aspects.

DATA MIGRATION. Often, organizations will need to migrate existing on-premises data to the cloud as part of their cloud strategy. Simple lift-and-shift migration of data from on-premises to cloud platforms can work, but in many cases, moving data is far more complicated. It may involve tweaking data models and interfaces for maximum performance on the new platform. This is best done in phases.

Additionally, if you're migrating data to the cloud, take the opportunity to improve it. Problems with data quality, data modeling, and metadata should be remediated before or during migration.

TOOLING. It is also important to consider the kinds of tools that will be used in the new environment. Aside from machine learning and AI analytics tools, many organizations want to automate parts of the analytics life cycle. This should be part of the plan. At TDWI, we're seeing interest in automated tooling in a number of areas. This includes automation of mundane tasks such as data cleansing.

Vendors are providing tools that include machine learning or other advanced algorithms to automatically detect potential data problems. Quality data is, of course, critical for data science. The old adage "garbage-in, garbage-out" applies here. Other tools utilize the same technology to identify sensitive data. Some tools will automate feature engineering or even build predictive models (more on this below).

It is a good idea as you make the move to data science to consider the kinds of tools you will want to use in your cloud deployment. Then ensure that the cloud environment supports multiple data and analytics tools (both open source and commercial) and that these tools are part of the cloud platform.

It can be useful to revisit historically reliable technologies and do a deep dive to determine if they are truly able to make the leap to the cloud. Consider how tightly they are integrated with the cloud platform, how they make use of cloud features, and how easy it is to use the tooling in a cloud environment.

SKILL SETS. Skills are often cited as a top challenge for organizations looking to implement data science. That means your plan will need to include how your organization will address both data science skills for building models and skills for putting models into production in the cloud.

Some organizations will look to low-code/no-code interfaces and augmented intelligence to help business analysts build machine learning models. This can work, but doesn't replace the need for those users to understand the algorithms behind the tooling. Some organizations will hire at least some data scientists to help train business analysts.

Additionally, organizations need to make sure a group (it might be called MLOps, DevOps, or DataOps) is in place to handle putting models into production, especially if they will use microservices and container technologies. Finally, there will be new IT skills required to utilize cloud-native environments.

GOVERNANCE. Data governance includes the rules and policies that are put in place to make sure data can be trusted and is in compliance. Data governance doesn't end when an organization moves to the cloud. In fact, in multiplatform environments, data governance becomes even

CREATE THE PLAN CONTINUED

more critical. This needs to be part of the plan. Additionally, for data science, not only data has to be governed—models will need to be governed as well. This will include policies for making sure the right processes are in place for versioning, validation, monitoring, and retraining.

VENDOR PARTNERSHIPS. Many organizations are tired of managing a wide and vast scope of vendors. Getting started can be slow. As part of the plan, your organization should consider how the cloud provider utilizes vendor partners as they supply the tools you need.

Vendors have been pivoting around cloud architectures. Certain vendor partners cover multiple areas and manage the full stack (infrastructure, software, data) so you can focus on your use cases. As mentioned, make sure vendors have tight partnerships that include technical partnerships with other vendors.

CONCLUDING THOUGHTS

Data science is the future of analytics, and the cloud provides an important platform for machine learning and AI because of its scalability, flexibility, and extensibility. As organizations mature in their analytics journey, TDWI routinely sees them moving to the cloud for both data management and analytics.

As part of this move, they will need to consider a number of factors including modern architecture and operationalizing the models they build, as well as planning for new skills, governance, and advanced analytics tooling that is tightly integrated as part of a cloud-native environment.

ABOUT OUR SPONSOR

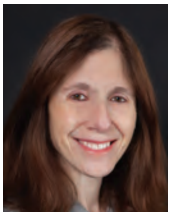
sas

As a leader in analytics, SAS helps organizations solve their toughest data challenges and make more intelligent decisions. Now, by migrating SAS analytics workloads to the cloud, organizations have greater flexibility and scalability to act on insights and drive relevant change. SAS takes on the design and delivery of software, infrastructure, and services in a managed environment—enabling users to sign up, log in, and get to work. By moving to a cloud-native implementation, SAS takes full advantage of the cloud's scalability, flexibility, and cost savings to help organizations achieve faster value and unlock new business opportunities. Learn more at sas.com/cloud.

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.

ABOUT THE AUTHOR



Fern Halper, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, text and social media analysis, machine learning, AI, cognitive computing and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her by email (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

ABOUT TDWI RESEARCH

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