
ORACLE® + DATASCIENCE.COM

# DataOps: An Agile Methodology for Data-Driven Organizations

**Crystal Valentine**, PhD, VP Technology Strategy, MapR

**William Merchan**, Chief Strategy Officer, DataScience.com

# Data Science and Machine Learning in the Enterprise

The promise of leveraging data for competitive advantage looms large in the modern enterprise. Organizations are collecting and curating data on a grand scale in support of cutting-edge, data-intensive applications both to increase revenues and improve operational efficiencies. As a result, data and algorithms can rightly be considered part of an organization's proprietary competitive edge.

The widespread adoption of big data computing platforms and commodity storage has sparked a renaissance of enterprise data science applications, including machine learning, deep learning, and artificial intelligence, that require large volumes of raw training data. The insights and operational efficiencies that can be gained through data science are among the most disruptive of modern enterprise applications. "Artificial intelligence is the apex technology of the information era," according to a Goldman Sachs Global Investment Research report.[1] "Advances in machine learning and deep learning have combined with more powerful computing and an ever-expanding pool of data to bring AI within reach of companies across industries. The ability to leverage AI will become a defining attribute of competitive advantage for companies in coming years."

Data science and machine learning have been used extensively for a set of well-known applications for decades, including quantitative finance, logistics, operations research, and advertising. Today, data science is becoming part and parcel of a much broader set of applications across all industries. For example, deep learning models make autonomous driving vehicles possible, manufacturers now use predictive analytics to avoid machine failure and improve factory output, and fraud prevention, targeted smart banners on websites, and recommendation engines are all built using machine learning.

The confluence of several secular trends have made possible today's renaissance of enterprise data science and machine learning practices. In particular, the emergence of mainstream enterprise-grade big data platforms and the commoditization of compute and storage servers, including public cloud infrastructure, have lowered the cost of managing and analyzing petabyte-scale data and have contributed to the widespread adoption of data science. These technological advancements have been accompanied by a concomitant focus on data and computing as a core business strategy.

Corporate structures have also changed to reflect this emerging priority. According to a 2017 New Vantage Partners survey[2], 55.9% of F1000 firms reported naming a chief data officer, up from only 12% in 2012. Also as of 2017, 69.4% of F1000 firms reported they were trying to create a data-driven culture, with 80.7% of executives declaring their big data investments successful; and 21% declaring big data to have been disruptive or transformational for their organizations.

In short, data science is emerging as a rapidly growing practice within the enterprise that is attracting focus and investment across industries. The technology community is responding by introducing new tools and platforms that can support data science workflows and the management of the large data sets required to train data science models.

[1] "Artificial Intelligence: AI, Machine Learning and Data Fuel the Future of Productivity," The Goldman Sachs Group, Inc., November 14, 2016.

[2] "Big Data Business Impact: Achieving Business Results through Innovation and Disruption," New Vantage Partners. 2017. http://newvantage.com/wp-content/uploads/2017/01/Big-Data-Executive-Survey-2017-Executive-Summary.pdf

# The Challenges of Putting Data Science Models into Production

While enterprise companies are making increasingly large investments in data science applications, many of them still struggle to realize the value of those efforts. Overwhelmingly, the common challenge of enterprise data science endeavors is that data scientists spend the majority of their time outside of their core competencies, working on data engineering and creating "glue" code to put together data pipelines. Developers at Google remarked, "It may be surprising to the academic community to know that only a tiny fraction of the code in many ML systems is actually devoted to learning or prediction … much of the remainder may be described as 'plumbing.'"[3]

According to Forrester Consulting, only 22% of enterprise companies are currently seeing a significant return from data science expenditures. Top performers among the more than 200 decision makers in business and customer insights, data science, and data engineering roles the firm surveyed last year had analytics budgets twice as large as their less-successful peers, as well as defined data science roadmaps. But those were not the only differentiating factors: 58% of data-driven companies reported that the ability to automate model deployment and create scalable APIs was critical to data science success — 24% higher than reported by average performers.[4]

A *deployed model* is a unit of code that is seamlessly integrated into a production environment and returns outputs based on the inputs it receives. For example, the model powering product recommendations on a retailer's website should examine a user's interactions on the site and then serve up relevant items for that user. Netflix's global recommendation engine, put into production in December 2016, saves the streaming service an estimated $1 billion a year by reducing monthly customer churn — in other words, it keeps viewers paying for their subscriptions by recommending shows they are more likely to watch. One of the attractive characteristics of data science models is that they can be improved and refined over time as new data is collected and new observations are made about how the model performs in production. However, the challenge of evaluating the performance of deployed models and rescoring and redeploying them is a logistical challenge. Netflix has spent more than a decade building its recommendation system; even so, its product team believes content recommendations could still be improved via faster and more effective A/B testing of deployed models.[5]

Netflix is far ahead of its peers. Only 11% of companies — and 13% of data-driven ones — have the capabilities required to iterate on models currently in production in order to improve their performance.[6] In many cases, even the process of getting a model off of a data scientist's laptop and into an environment where it can power an application is arduous. Traditionally, a data scientist must hand over his or her model code — typically written in R or Python — to the engineering team to be refactored and rewritten into a production stack language that is more attuned to scalable application building, such as Java or C++. Only then can it be moved into an engineering production environment for testing and fine tuning, and eventually, a full rollout. This entire pipeline requires extensive support from IT, which manages the underlying systems and resources needed to scale the application.

---

[3] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, and J.-F. Crespo. Hidden technical debt in machine learning systems. In Neural Information Processing Systems (NIPS). 2015. https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.

[4] Forrester, "Data Science Platforms Help Companies Turn Data Into Business Value," https://cdn2.hubspot.net/hubfs/532045/Forrester-white-paper-data-science-platforms-deliver-value.pdf

[5] Netflix, "The Netflix Recommender System: Algorithms, Business Value, and Innovation," http://delivery.acm.org/10.1145/2850000/2843948/a13-gomez-uribe.pdf

[6] Forrester, "Data Science Platforms Help Companies Turn Data Into Business Value."

Just three years ago, the software analytics firm New Relic found that a quarter of companies were only deploying code on a monthly basis, while 13% didn't plan to deploy any code for an entire year. But as businesses strive for faster time to value, that paradigm is changing; in 2015, 57% of companies were releasing code into production weekly, multiple times per week, or even multiple times per day, and 66% planned to maintain or reach that cadence by 2016.[7] (Here's an even more extreme example: Amazon is reportedly deploying code every 11.7 seconds.)[8] This shift is indicative of a larger trend in the way companies are approaching data-related tasks: Instead of waiting to deploy finished projects at the end of a long waterfall-style production cycle, companies are taking an agile approach that allows for smaller, faster rollouts.

Automation and scalable APIs are both helping to foster this change. To continue using our recommender system example, retail giant Walmart makes its product recommendation available via an API, which developers can use to power product recommendation widgets in marketing emails.[9] Now, the same idea is being applied to internal data science pipelines at enterprise companies. The data scientist makes his or her model available via an API, and the engineering team can plug it in where appropriate, eliminating the need to refactor or rewrite code. The company's systems – business intelligence tools, call center software, or its public-facing website, for example – can "talk" to the API and receive the model's outputs continuously.

Agile companies like Stitch Fix are embracing a work flow where data scientists can "deploy their ideas with autonomy" using services and frameworks maintained by engineers, rather than relying on engineers to do the bulk of the production work.[10] But many others businesses are still struggling to embrace a faster, more agile approach to leveraging big data. While there are many factors at play, one of the most critical is a lack of a unified, secure platform for data science and analytics work.

[7] New Relic, "Going to Market Faster: Most Companies Are Deploying Code Weekly, Daily, or Hourly," https://blog.newrelic.com/2016/02/04/data-culture-survey-results-faster-deployment/

[8] TechBeacon, "10 Companies Killing it at DevOps," https://techbeacon.com/10-companies-killing-it-devops

[9] Imart, "Product Recommendation API," https://developer.walmartlabs.com/docs/read/Product_Recommendation_API

[10] Multithreaded, "Engineers Shouldn't Write ETL: A Guide to Building a High Functioning Data Science Department," http://multithreaded.stitchfix.com/blog/2016/03/16/engineers-shouldnt-write-etl/

# A Platform Approach

Unlike the traditional waterfall model of product development, an agile approach to data science model deployment embraces collaboration and creativity. Changes can be made as a project evolves, and stakeholder feedback is ongoing. This is difficult to achieve in an environment where work is siloed, data is inaccessible, and data scientists lack ownership of their projects.

A winning data strategy brings all elements of big data analysis together and makes data science work visible, shareable, reproducible, and standardized. As described above, providing some elements of self-service — such as the ability for data scientists to deploy models as APIs or securely access the data sets they need — can streamline the data science process. But before model deployment occurs, there are several other areas that need to be managed effectively, a task that is often made easier with a holistic platform approach to data science and machine learning.

## Environment Management

Data scientists need somewhere to build models, and in many cases, that place is on their local machines. In this common scenario, the code a data scientist writes may not behave as desired — or work at all — when run in a new environment that contains package versions or toolsets that are inconsistent with what the data scientist originally used.

To avoid this problem, some companies task their IT teams with building environments for data scientists on an ad hoc basis. Agile companies, however, take a less burdensome — and more self-service — approach. With the help of a containers-as-a-service platform like Docker, it becomes a relatively simple task for your IT team to create an ideal data science environment, equipped with whatever tools — either open-source or proprietary — your data scientists prefer. Your data scientists can then create as many environments as they need from that base Docker image. In fact, Docker claims that businesses using its containers improve the speed at which they deploy code seven times over.[11]

## Resource Management

IT teams are also commonly tasked with managing the resources needed by data scientists to run models and build analyses. If your data science team is hogging resources it doesn't need, that can drive up costs; it's important that you have guardrails in place. Providing your IT team with a platform that supports multi-tenancy can provide optimal resource utilization and the ability to segregate test data and processing from production workloads, if desired. Moreover, a platform with self-tuning and self-healing cluster management capabilities and a centralized administrative dashboard reduces administrative overhead and makes it easy to scale resources dynamically.

## Data Engineering and Data Curation

Traditionally, companies have regarded raw data as something to be extracted, transformed, and loaded into data warehouses before it can be accessed by an analyst. But this approach is neither agile nor conducive to good data science; instead, the most advanced companies are leveraging enterprise-grade big data platforms for a cost-effective, flexible, and fast way to access and secure multiple data types – including raw files, event streams, and document tables – supporting a self-service "data marketplace" access paradigm with less reliance on IT and administrators.

## Moving Toward DataOps

When brought together in a platform, all of these elements can help create a more agile, creative, and collaborative environment for doing data science work at your organization. But we've only scratched the surface of what it means to implement a  DataOps workflow.

[5] Docker, "Company," https://www.docker.com/company

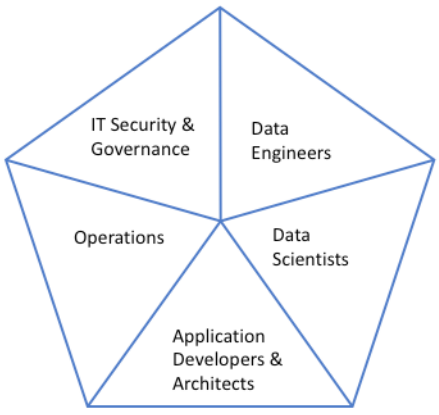
# DataOps

DataOps is an agile methodology for developing and deploying data-intensive applications, including data science and machine learning. A DataOps workflow supports cross-functional collaboration and fast time to value. With an emphasis on both people and process, as well as the empowering platform technologies that underlie it, a DataOps process allows each collaborating group to increase productivity by focusing on their core competencies while enabling an agile, iterative workflow.

## Cross-Functional Collaboration

A DataOps workflow paradigm must be embraced and implemented by several different functional groups within an organization which all collaborate to deliver business value:

- Data engineers aggregate and curate the data sets that are used for training and evaluating data science models.
- Data scientists build and publish data science and machine learning models, often through experimentation and iteration. They focus on developing rigorous and efficient models, leveraging their favorite statistical languages, such as Python or R, as well as frameworks for developing machine learning and deep learning models, such as Spark, TensorFlow, Caffe, and Theano.
- Application developers and architects build end-to-end applications that include the models developed by data scientists in the application logic.
- Data governance and IT security define the data access controls in support of a self-service "data marketplace." This enables data scientists to access historical data for model training and benchmarking while production data sets may be accessed in a read-only manner for model tuning and refining.
- The operations team, or sometimes DevOps, deploys applications to the production environment to support SLAs.




A DataOps methodology requires cross-functional collaboration.

## A Platform for DataOps

While DataOps focuses largely on people and process, it also requires an enterprise-grade platform to enable collaboration and the sharing of data and compute resources by the different groups involved. Rather than having each group work off of siloed technology and data, a collaborative effort should be able to leverage a single, unified data platform. Using a single platform is key to agility, reduces the need to copy or move large data sets, and supports a holistic approach to data access and security.

The technical requirements of a platform that can support a DataOps process include:

- Enterprise-grade reliability. The platform should be considered mission-critical, so it should feature native replication and snapshots for high availability and disaster recovery.
- Native support for any data type to accommodate diverse and evolving data sources. All data, including files, tables, and event streams should accessible in a single global namespace.
- Linear and unlimited scalability to grow with your business and use cases.
- Support for distributed architectures. The platform should feature the ability to manage data across different physical locations, including public and private clouds as well as edge devices.
- Full-featured support for data science model development with built-in collaboration tools, including tools for model sharing and versioning and report building.
- Support for all the popular data science languages and tools.
- Support for a "model publication" process in which data scientists can push trained models to the application development and operations teams so they can be leveraged by production systems.
- Multitenancy. The platform should support application development, data science, and production workloads with native support for multiple compute engines as well as resource allocation and prioritization/scheduling tools.
- Self-service data access through a metadata-driven data marketplace. The platform should empower the security and governance teams to enforce privacy and security policies with granular access control expressions while promoting a self-service, agile data access workflow.

## Benefits

A DataOps methodology yields many benefits to data-driven organizations, but principle among them are:

- Agility. Data scientists can iterate rapidly to improve models. The publication of new models can happen independently of the application development work, and new models can be deployed to production without interrupting the production application's operation.
- Increased productivity. Each group can focus on their core competencies. Data scientists do not get bogged down in doing the "plumbing" work of finding, copying, curating, and transforming data. Application developers do not waste time refactoring code written by data scientists so it can run in production.
- Security. With a unified data platform, organizational data access and privacy policies can be enforced holistically across organizations. Model development and application deployment activities inherit from the data access policies specified by the governance group.

# Conclusion

The mantra of the Big Data 1.0 movement in the mid-2000s was that aggregating large volumes of data was itself valuable. We now better appreciate that aggregating data into a lake is not enough; data does not bring value until it is leveraged to impact either top-line growth or profitability through increased operational efficiency. One of the most impactful ways of operationalizing data is through the use of data science and machine learning to build intelligent applications.

A DataOps methodology focuses on improving the business value of data science and machine learning investments by speeding time to market for intelligent applications. A DataOps practice makes data consumable in an efficient and agile way while still respecting governance and security policies.

While DataOps is still emerging as an enterprise practice for organizing the work of small teams involved in a collaborative process to build data applications, it represents a significant new trend. The recognition that data-intensive applications have their own set of considerations when it comes to managing and securing large, complex data sets while still enabling agile access to that data by the people who need it is a paradigm shift. As data access is often the bottleneck in supporting data-intensive applications, an intelligent application development practice should consider the use and management of data as an organizing principle. Thinking about the management and access to data as part of the development process turns conventional application-centric thinking on its head. New DataOps development practices are a sign that we are moving toward a new reality in which IT is no longer a bottleneck to productivity but instead part of a fluid process that emphasizes self-service and faster time to market while still being enterprise grade.

ORACLE + DATASCIENCE.COM

Connect with us on social media:

@DataScienceInc

https://www.linkedin.com/company/datascienceinc

@DataScienceInc

https://www.facebook.com/datascience

Oracle Corporation, World Headquarters 500 Oracle Parkway Redwood Shores, CA 94065 USA