Considerations and Approaches for Scaling DataOps

Hint: Modeling It on DevOps Won't Work

April | 2021

Table of Contents

Executive Summary

To fully harness data at the level that modern competitiveness requires, a business needs to become a cognitive enterprise that builds data-driven insight into all its operations. That transformation requires scalable, repeatable processes for managing data pipelines and their output, with quality controls built in. Put another way, organizations need a formal DataOps program instead of taking an ad hoc approach to running data management, analytics, big data and other BI programs. Unlike departmental initiatives or one-off data projects, DataOps requires changes that reach further across organizational culture and information architecture. Such changes are hard to make; 92 percent of companies say culture is their leading impediment to becoming a data-driven organization[1].

This thought paper will highlight the unique DataOps characteristics, explain the most important differences between DevOps and DataOps, identify the relevant DataOps challenges, and provide enterprise guidance for establishing and scaling DataOps programs.

Unlike departmental initiatives or one-off data projects, DataOps requires changes that reach further across organizational culture and information architecture.

Organizations need a formal DataOps program instead of taking an ad hoc approach to running data management, analytics, big data and other BI programs.

DataOps is becoming increasingly important to enterprise competitiveness, but it is hard to start and even harder to scale. There is a tendency to model DataOps efforts after DevOps, which most organizations now have some experience with. This approach is problematic, and the most serious problems tend to surface when organizations try to scale their DataOps efforts. And programs will need to scale – data and the demand for data-driven insights are both growing quickly. The average company was managing 5,000 datasets in 2020, up from 4,300 in 2018, a 16 percent increase[2].
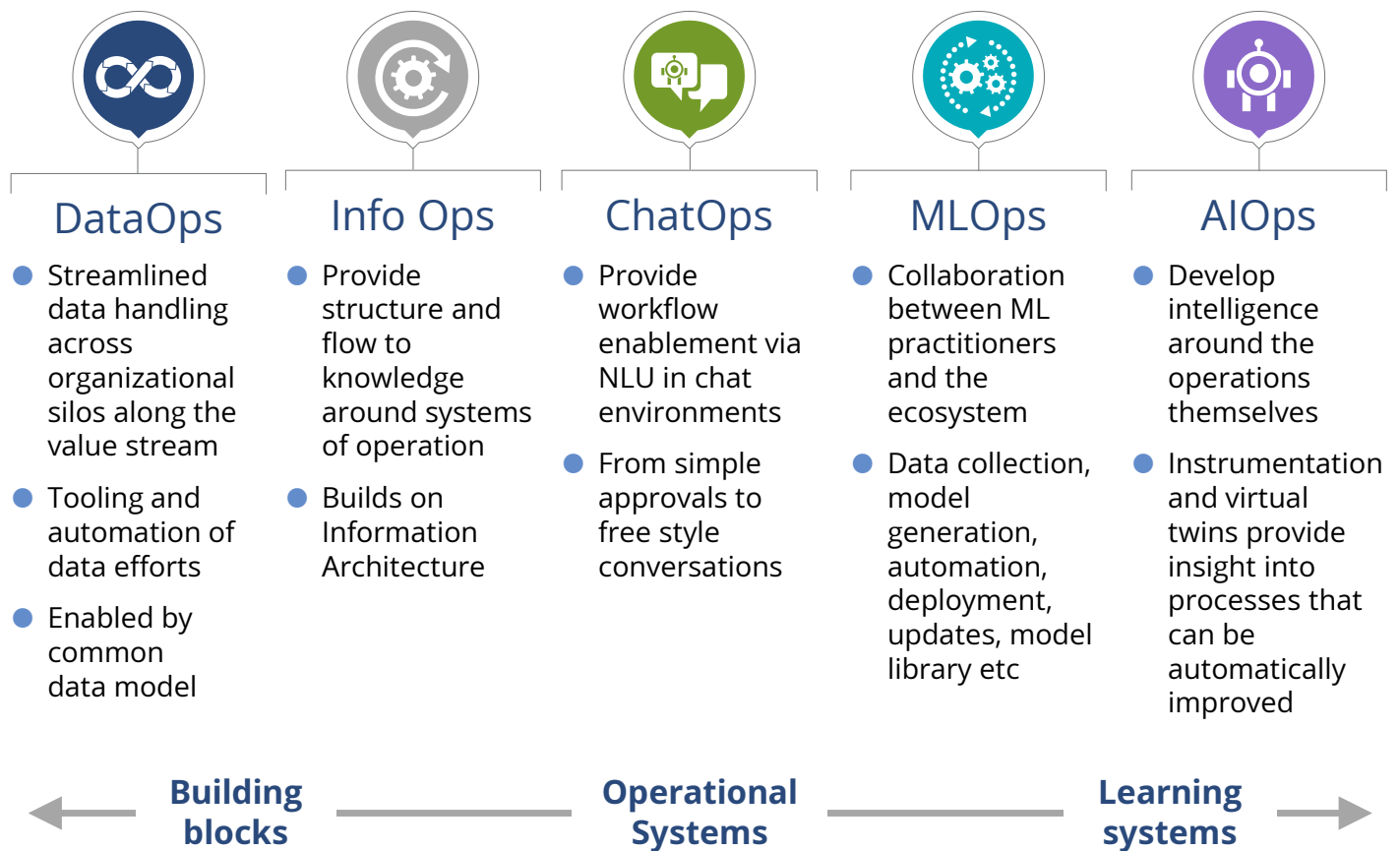
What is DataOps?

DataOps covers the entire life cycle for data, including data acquisition, categorization, cleansing, correlation and enrichment, data analytics (including model building), data preparation and pipeline management, QA, visualization, and reporting. It includes elements of DevOps (to promote speed and a culture of shared responsibility) and statistical process control (for continuous quality and to enhance development speed). It also involves "platformation" of unseen future value (to advance the organization to a platform-based delivery model for data insights whose consumers and use cases are not readily evident up front).

To understand what DataOps is, it is useful to consider what it is not. DataOps is distinct from InfoOps, AIOps, MLOps or ChatOps, which are all separate practices with different levels of maturity and adoption. The graphic below provides capsule summaries of these different practices.

More importantly, DataOps is not DevOps. DataOps is often considered "DevOps for data" but this is an incomplete, erroneous view that can lead to problems with DataOps program design, management, output and expectations.

DataOps
- Streamlined data handling across organizational silos along the value stream
- Tooling and automation of data efforts
- Enabled by common data model

Info Ops
- Provide structure and flow to knowledge around systems of operation
- Builds on Information Architecture

ChatOps
- Provide workflow enablement via NLU in chat environments
- From simple approvals to free style conversations

MLOps
- Collaboration between ML practitioners and the ecosystem
- Data collection, model generation, automation, deployment, updates, model library etc

AIOps
- Develop intelligence around the operations themselves
- Instrumentation and virtual twins provide insight into processes that can be automatically improved

← **Building blocks** — **Operational Systems** — **Learning systems** →

Source: ISG

Figure 1 – Capsule View of the XOps Waterfront

Why DataOps Needs Dedicated, Tailored Management

At a high level, both DataOps and DevOps make information and services available to business users. They share fundamental goals and both rely on current information systems and sources. However, when it comes to effectively managing these operations, their differences are more important than their similarities. Here are a few key differences that make different program management approaches necessary:

- DevOps manages code and DataOps manages data. Data is inherently more complex than code. It is created in the real world, is produced organically, and can't be as tightly controlled. In DevOps, code is created by developers, who have much more control over inputs and outputs than data scientists.

- While DevOps focuses on the flow of features and availability on production environments as success factors, DataOps needs to address the quality of the path as a deliverable/achievement as well as the goals themselves (data quality in production).

Compared with DevOps, however, the DataOps talent shortage is more acute, the field is more nascent and the roles themselves are more expensive.

These and other nuances make DataOps harder to establish and scale. Here are additional factors that impact the success of DataOps initiatives.

There has been an explosion in data volume and in the types of data sources that enterprises are seeking to harness, especially as they bring more unstructured data into their environments. This growth increases DataOps complexity, especially when it comes to scaling.

Talent Issues

DevOps and DataOps both suffer from a shortage of qualified talent (service providers have the same problems as enterprises in hiring and retaining specialists, which is something to check into when evaluating providers). Compared with DevOps, however, the DataOps talent shortage is more acute, the field is more nascent and the roles themselves are more expensive.

Complexity

DataOps is expected to deliver data pipeline quality and management. There is no equivalent step in the DevOps process. There has been an explosion in data volume and in the types of data sources that enterprises are seeking to harness, especially as they bring more unstructured data into their environments. This growth increases DataOps complexity, especially when it comes to scaling.

As for process, in DevOps there is a certain linearity to the progression from requirements to delivered functionality. DataOps typically deals with convolutions in information cycles and inputs of unknown provenance.

There is rich variety in the DataOps playground, from plain vanilla data pipelines to actual ML models, to updated predictive models/business weightages/ conversation trees/application heuristics – all of which need to be monitored and adjusted. Estimating the effort needed to satisfy DevOps asks (for example, requests for new features, environments or configurations) is easier than estimating DataOps asks (which typically might be for new models or to find validation data to test a new hypothesis). The feasibility and availability of raw material is often in question. Even when needed data from multiple sources are available, combining them may not be meaningful or possible.

The average organization is managing more than 5,000 datasets. Because of the complexities in source data and processes, data professionals are only spending 14 percent of their time on actual analysis, and much more on data wrangling[3]. Automation can reduce data wrangling time, but it is only one piece of a comprehensive DataOps management process.

Automation can reduce data wrangling time, but it is only one piece of a comprehensive DataOps management process.

Quality Management

It is easier to maintain quality and consistency for code than it is for data. Testing and QA are more mature for DevOps than DataOps. In addition, data quality tends to decline as data pools grow.

Hopefully, it hasn't happened at your enterprise, but many analytics programs have been undermined because the organization couldn't get a handle on master data management and other data quality management tasks. For example, 40 percent of organizations reported their staff doesn't trust the data insights the enterprise produces, and 78 percent said data debt is a problem for their organizations[4]. Underlying data quality issues such as these surface in DataOps, where data quality and consistency are recognized as being key to success.

In addition, data quality tends to decline as data pools grow. Hopefully, it hasn't happened at your enterprise, but many analytics programs have been undermined because the organization couldn't get a handle on master data management and other data quality management tasks.

Guidance for Creating and Managing a Scalable DataOps Program

Clearly, organizations need to accept from the beginning that DataOps must be envisioned and run differently than DevOps. Establishing this belief and the culture to support it are important to success. Culture was identified as the leading impediment to becoming a data-driven organization in 2021[5].

> In addition to culture, a successful DataOps program needs to address specifics related to data, the process for developing services and overall management.

How **can** you account for the differences and create a program with the proper structure, staffing, practices and governance? In addition to culture, a successful DataOps program needs to address specifics related to data, the process for developing services and overall management. ISG presents the following four guiding principles and some suggested approaches for putting them into practice.

- **Describable data and processes.** There must be clarity about what each element and process needs, and how each modifies the data. To do this, it is critical to distill business and domain characteristics at the lowest level. If elements can be described, it is a good sign that processes can be repeatable, which is essential for scalability.
- **Liveness and correctness.** Trust results and insights from fresh data sources and the most recently used insight generators more than from previous ones. A corollary to that is to cull data flows that are not being used or adopted. Ensure that each data element and flow is correct in its environment by eliciting feedback at the process level and the consumption (user) level.

- **Enterprise understanding.** Change management is crucial to DataOps program success. Business users need to have confidence in analytics results, and they need to be able to act on them. That requires education and trust building (as noted, 40 percent of organizations reported lacking internal trust in the enterprise's data insights). The personnel that serve as the interface between the data analytics team and enterprise business usersplay a critical role. Focus on hiringand training talent who can both communicate the business user needs to the DataOps team and makeDataOps possibilities, limitations and requirements clear to business users.

> It may be tempting to initially bypass corporate IT to start a project quickly and prove the concept early. However, to scale DataOps you will need to heavily involve IT.

- **Speed.** DataOps is not a "set it and forget it" function because its processes will frequently need to be tweaked to accommodate new input and output requests. Therefore, IT must not be allowed to stagnate. It may be tempting to initially bypass corporate IT to start a project quickly and prove the concept early. However, to scale DataOps you will need to heavily involve IT.

Build these principles into the following foundational steps for a DataOps program.

Create a DataOps or Analytics Center of Excellence

Having a CoE helps bring order to unchecked individual business unit initiatives. Centralizing learning, development and data services production helps organizations develop best practices and make sure they are consistently applied. When a CoE is available, business units will be less likely (or even prohibited) from acquiring and learning their own tools. This not only saves in maintenance and support costs, it supports scalability because data management tools and platforms are often incompatible, which restricts the ability of data to move across systems. An incompatible toolset also limits collaboration and integration opportunities. Scalability and compatibility limitations might not be apparent in pilot projects or department-level initiatives, but they have become true problems for many enterprises.

Pursue Automation Where Possible

Scaling will require automation to handle the increase in available data. Requests for new data outputs and services will also grow, which means the development process must also be scalable. Many DataOps tasks can be automated to help mitigate all the leading challenges, including the lack of skilled resources, the amount of time these professionals must spend on data wrangling vs. data analysis, explosive data volume growth and quality control.

There are two strong ways to infuse automation into DataOps. First, help the DataOps team by implementing a DataOps-specific platform and tools (which can be purchased or built). If your team does not have expertise in choosing and implementing platforms, you can get expertise from the market to get you started. Depending on the solution, data platforms can provide many automated data management, testing, QC and development functions out of the box.

Later, open the platform up to limited self-service so business users can do their own analysis and create their own visualizations. Data democratization is currently a strong trend, and data self-service is a high priority for enterprise investment. Enabling self-service requires collaboration with IT to give users access to the enterprise data warehouse and other data sources while ensuring compliance and adherence to policies. It also requires a data platform with an interface that works well for non-specialist users.

> There are two strong ways to infuse automation into DataOps. First, help the DataOps team by implementing a DataOps-specific platform and tools…later, open the platform up to limited self-service so business users can do their own analysis and create their own visualizations.

Develop and Nurture DataOps Culture

Gaining organizational support from the top down (executive level) and bottom up (operations and business user level) makes DataOps more effective and easier to scale. You can nurture a DataOps-friendly culture by communicating what is possible with DataOps, what is being done now and how the enterprise is benefiting. This will help employees see themselves as both consumers and producers of data, which will lead them to become better data stewards.

Part of building a strong culture is providing formal training to improve technology and data literacy throughout the organization. The staff development program should also include grooming personnel to be data champions and data stewards. Ideally, staff in business units will gravitate to these roles so it isn't all left to the DataOps team. Having trained staff and a supportive culture throughout the organization will also be a big help to governance, which is often thorny for DataOps.

> Part of building a strong culture is providing formal training to improve technology and data literacy throughout the organization.

Create a DataOps-specific Governance Program

By now it should be clear why putting DataOps governance under a DevOps or DevSecOps umbrella isn't appropriate. There is growing recognition for the need and value to manage DataOps specifically, which is why spending on data governance is projected to increase by 271 percent to $5.7 Billion between 2020 and 2025, according to one study[6].

One reason DataOps governance is important and specialized is because the risk of non-compliance escalates significantly when data is combined with additional sources and used in new ways. For example, Google's $57 Million fine in 2019 for GDPR violations could be viewed as a data governance failure because regulators cited the company for not obtaining user consent for each distinct data processing operation the company executed. As different models repeatedly reuse the same data points, risk increases, often in ways that were not foreseen when the data was collected.

Compliance aside, the fundamental activity of DataOps is different than that of DevOps or providing general IT service, so the typical metrics and governance methods for those functions are not especially relevant. At a minimum, DataOps needs its own metrics. Examples include:

- Number of data artifacts
- Number of users for data artifacts
- Data errors
- Schema changes
- Ratio of data engineers to data scientists
- Number of self-service users

This is not a comprehensive list. It is included to illustrate that DataOps governance requires new thinking.

Assess What is Best Done In-house and What Should be Sourced

DataOps initiatives are uncharted territory for many companies. It is a critical capability that will enable efficiency, foster differentiation and drive successful business outcomes.

So a fine balance has to be struck between leveraging external expertise for expediency and building in-house strengths in a core area. A scaffolding approach offers a model for successful adoption. Mature organizations that are at steady state typically exhibit the division of responsibilities shown in Figure 2.
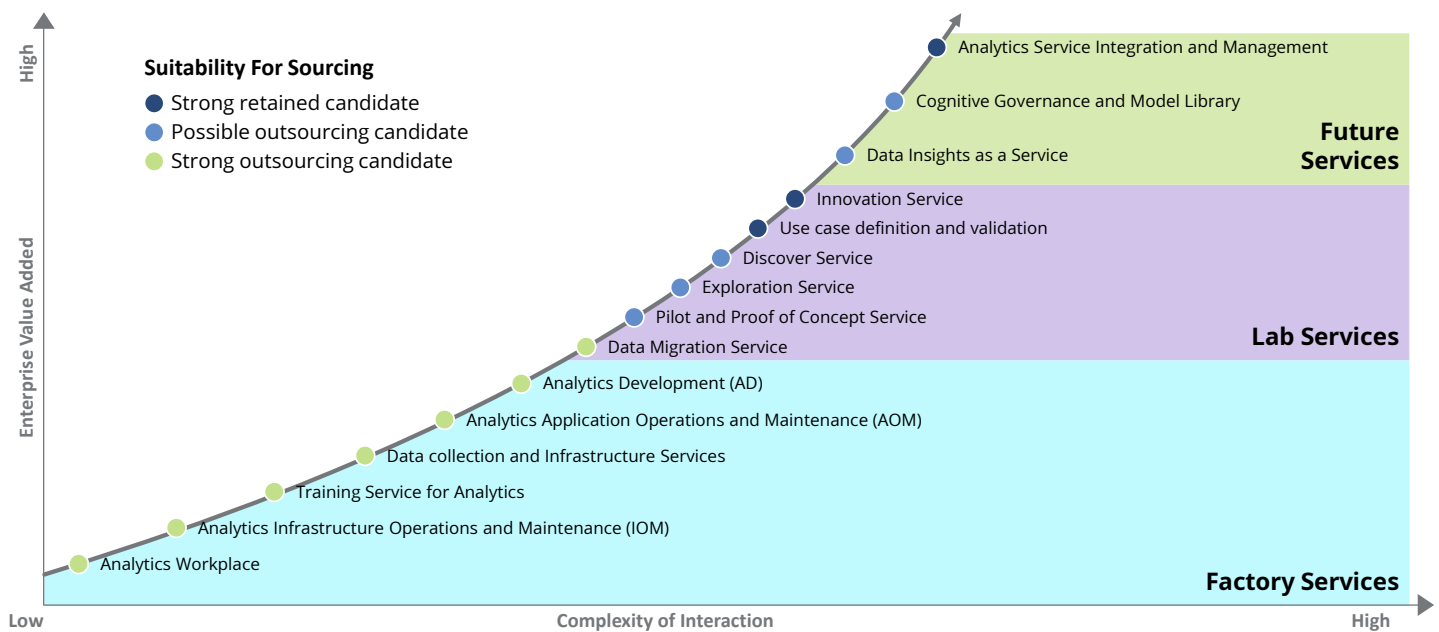
Source: ISG

Figure 2 – Sourcing Decision Swoop Chart for Analytics Service Catalog

Conclusion

The tools for collecting, combining and analyzing data are exploding, and demand for information and insights from it might be growing even more. Enterprises need to take a comprehensive approach to managing data and internal demand. If not, they risk becoming bogged down and left behind while continually dealing with ad hoc requests and trying to get various departmental datasets and tools to work together.

Organizations need a formal DataOps program, and DataOps needs to be managed distinctly, with provisions for its specific requirements and risks. Treating DataOps as DevOps puts enterprises at risk of non-compliance and quality problems resulting from insufficient data management. DataOps requires its own tooling, metrics and culture. Contact ISG to learn more about how we can help your organization put an effective DataOps program in place.

References

[1] New Vantage Partners, "Big Data and AI Executive Survey 2021."

[2] Nexla, "2018 Definitive Data Operations Report."

[3] Nexla, "2018 Definitive Data Operations Report."

[4] Experian, "2020 Global Data Management Research Benchmark Report."

[5] New Vantage Partners, "Big Data and AI Executive Survey 2021."

[6] MarketsandMarkets, "Data Governance Market Worth $5.7 Billion by 2025," July 27, 2020.

Meet the Team

Shriram Natarajan
Director, Digital Strategy and Solutions

Shriram Natarajan leads ISG's Digital Solutions and Services in the Americas with a focus on data capabilities and maturities. He offers clients his 20+ years of experience in Technology Consulting and Adoption. His experience spans from leading core software product development to delivering IT services for application portfolios both onsite as well as offshore – which he combines this with digital trends to bring thought leadership to his engagements.

John Burnell
Author

John has been an ISG contributor since 2008. He is an author and editor for content at ISG Research and is heavily involved in IT market research. He applies the skills and lessons learned to support the next wave of disruptive technologies and processes, working in outsourcing, mobile & wireless computing, analytics, cloud computing, telehealth and other healthcare transformation, XaaS, data center automation, virtualization, automated data collection and more. Prior to joining ISG John was an award-winning journalist in the technology industry.

Aparna Gajanan
Senior Manager, Provider Services

Aparna has 13 years of experience researching and writing on the impact of digital transformation across industries. Her work has involved conducting in-depth research and analysis focused on identifying future trends and developing differentiated frameworks for clients. Prior to joining ISG, Aparna developed a robust research practice for digital and strategic insights at Accenture.

Acknowledgements

We would like to thank editor **Jack Kirshbaum** for his valuable insights and advice in preparation this thought paper.

ISG

imagine your future®

ISG (Information Services Group) (Nasdaq: III) is a leading global technology research and advisory firm. A trusted business partner to more than 700 clients, including more than 75 of the top 100 enterprises in the world, ISG is committed to helping corporations, public sector organizations, and service and technology providers achieve operational excellence and faster growth. The firm specializes in digital transformation services, including automation, cloud and data analytics; sourcing advisory; managed governance and risk services; network carrier services; strategy and operations design; change management; market intelligence and technology research and analysis. Founded in 2006, and based in Stamford, Conn., ISG employs more than 1,300 digital-ready professionals operating in more than 20 countries–a global team known for its innovative thinking, market influence, deep industry and technology expertise, and world-class research and analytical capabilities based on the industry's most comprehensive marketplace data.

www.isg-one.com