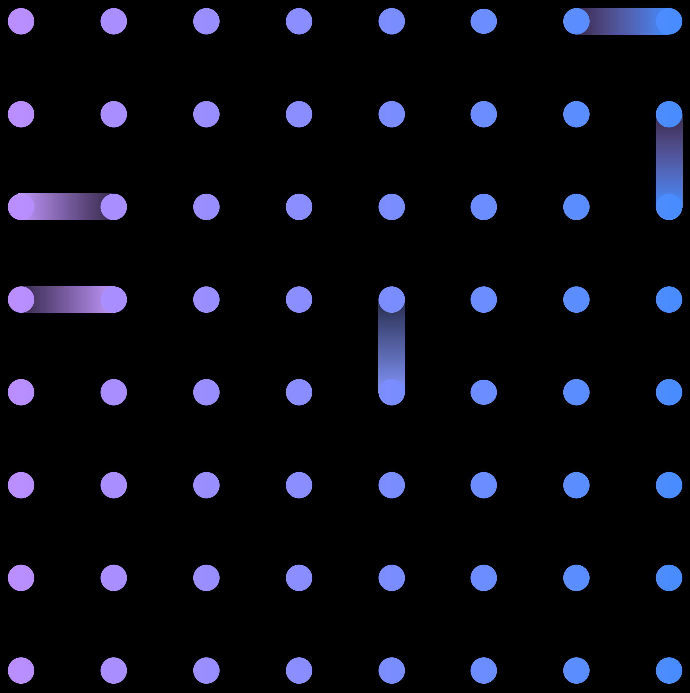
Deliver business-ready data fast with DataOps

An introduction to the IBM DataOps methodology and practice



Contents

Introduction 3

Defining DataOps 3

Comparing objectives 4

IBM DataOps program 7

The IBM Cloud Garage methodology meets DataOps 7

The impact of a successful DataOps practice 10

Conclusion 11

Appendix: DataOps Pilot Program Template 12

Highlights

- DataOps is the orchestration of people, process and technology to deliver trusted, high-quality data to data citizens fast
- Powered by automation, DataOps tackles the challenges associated with inefficiencies in accessing, preparing, integrating and making data available
- The IBM DataOps practice leverages extreme automation to drive significant and measurable impact on the following capabilities: data curation services, metadata management, data governance, master data management and self-service interaction
- IBM provides a path to a DataOps practice with a prescriptive methodology, artificial intelligence (AI)-enabled automation and the IBM DataOps Center of Excellence
- A DataOps workshop is an integral part of the DataOps roadmap, helping an organization to assess their DataOps maturity and to plan the execution of a pilot project

Introduction

Data is the fuel for innovation and sustaining a competitive advantage. It's the key ingredient for driving analytics and understanding business trends and opportunities. Unlocking the value of data in new ways can even accelerate an organization's journey to AI.

Yet, when data-related projects fail to deliver the promised return on investment (ROI), stakeholders want to know why. According to Experian's 2019 Global Data Management Research report, 89% of businesses report that they struggle with managing data. These struggles include delays in insight and a lack of trust in underlying data.[1]

Understanding an organization's business goals is critical to developing an effective data strategy for analytics and AI. For any business model to work, it must meet client needs. Success depends on streamlining data operations with an integrated business-ready data pipeline, which provides a complete and consistent view of the business at any point in time.

The expectation to achieve faster results is one that continues to rise. Businesses everywhere are looking for ways to improve their operational efficiency and effectiveness to enable the best decision-making, especially due to many silos within an organization. These two factors cause business leaders to seek new ways to tackle their biggest challenges within a single framework.

For organizations seeking a transformation within their data operations, automation technology can deliver a competitive advantage. Data becomes valuable when trusted business-ready data helps drive differentiated insights and operational excellence for organizations.

The purpose of this white paper is to highlight the benefits of the DataOps methodology, practice and roadmap.

Defining DataOps

Data operations (DataOps) is the orchestration of people, process and technology to deliver trusted, high-quality data to data citizens fast. The practice is focused on enabling collaboration across an organization to drive agility, speed and new data initiatives at scale. Using the power of automation, DataOps is designed to solve challenges associated with inefficiencies in accessing, preparing, integrating and making data available.

The potential benefits of DataOps include significant productivity gains in delivering information and data to individuals and improving processes to gain efficiency and optimization. Automated data operations that include AI data-led initiatives can help drive the following outcomes:

- Deliver integrated business-ready data that drives analytics and AI at scale
- Achieve operational efficiency
- Enable data privacy and compliance

89%

of businesses struggle with managing data.[1] Understanding an organization's business goals is critical to developing an effective data strategy for analytics and AI.

DataOps is not DevOps

The majority of organizations have implemented some level of DevOps within their development disciplines. The widespread familiarity of the DevOps practice and similarity in naming convention have invited comparisons with the emerging DataOps practice. While both are methodologies to drive operational best practices, they each have their unique place in an organization.

In the table below, see how both practices compare when it comes to objectives and benefits to an organization.

Comparing objectives

	DataOps	DevOps
Key focus	Business-ready trusted, high-quality data available for use fast.	Application and software development
Transformational objectives	– Fuel continuous and fast innovation for the business by enabling self-service access to trusted, high quality data for all data citizens – Enable continuous delivery of data by automating data governance, integration and while safeguarding regulatory concerns – Provide a feedback loop for continuous learning from all data citizens by monitoring and optimizing the data pipeline	– Speed continuous innovation of ideas by enabling collaborative development and testing across the value chain – Enable continuous delivery of these innovations by automating software delivery processes and eliminating waste—while still helping to address regulatory concerns – Provide a feedback loop for continuous learning from customers by monitoring and optimizing the software-driven innovation
Efficiency objectives	– Correct the misalignment of people and goals by fostering closer links between IT system support, operations and the business – Accelerate the delivery of changes and improve delivery quality by introducing automation throughout the data delivery cycle – Improve insight into the real value of metadata and data by using results to drive optimization	– Correct the misalignment of people and goals by fostering closer links between developers, operations and the business – Accelerate and remove errors from the delivery of changes by introducing automation throughout the development cycle – Improve insight into the real value of applications by using customer feedback to drive optimization

DataOps is the orchestration of people, process and technology and to sustain a commitment to a DataOps practice requires deep collaboration across all functions. It requires a focus on cultivating data management practices and processes that improve the speed and accuracy of analytics.

People and process

DataOps supports highly productive teams with automation technology to help deliver efficiency gains in both project outputs and time taken to deliver. However, to experience the benefits, the internal culture needs to evolve to truly be data-driven. With more business segments requiring and wanting to manage data to drive contextual insights, the time is right to do the following:

- Increase the quality and speed of data flowing to the organization.
- Obtain commitment from leadership to support and sustain a data-driven vision across the business.

This type of transformational change begins by understanding the true goals of the business. *How does data inform the decisions and services impacting customers? How can data help maintain a competitive advantage in the market? What are the financial priorities that data can help us solve?*

DataOps leaders should define the roles all data citizens play to drive the culture and the DataOps practice moving forward. Each organization has unique needs where stakeholders in IT, data science and the business lines need to add value to drive a successful business. Also, leveraging existing data governance committees and learnings from tenured data governance programs helps establish this culture and commitment as governance is one of the driving forces needed to support DataOps.

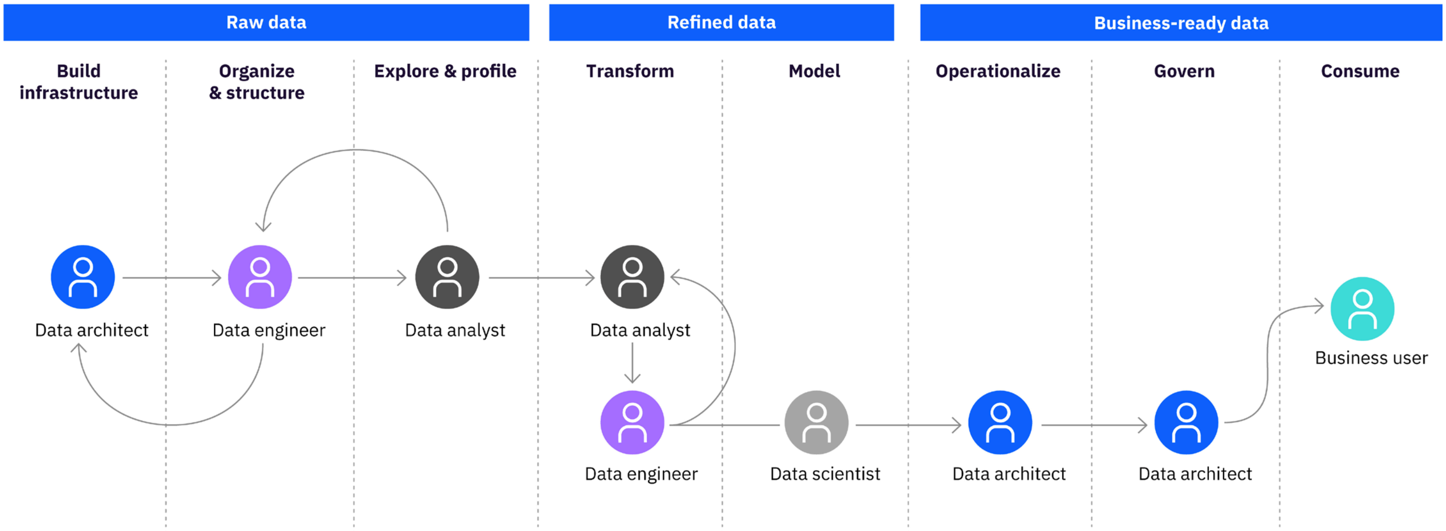


Figure 1: Example of a DataOps workflow by roles

Technology

At the core of DataOps is an organization's information architecture. *Do you know your data? Do you trust your data? Are you able to quickly detect errors? Can you make changes incrementally without "breaking" your entire data pipeline?* To answer these questions, the first step is to take inventory of your data governance and data integration tools and practices. Tooling is necessary to support any practice that relies on automation.

When considering tooling to support a DataOps practice within an organization, think about how automation in these five critical areas can transform a data pipeline:

1. Data curation services
2. Metadata management
3. Data governance
4. Master data management
5. Self-service interaction

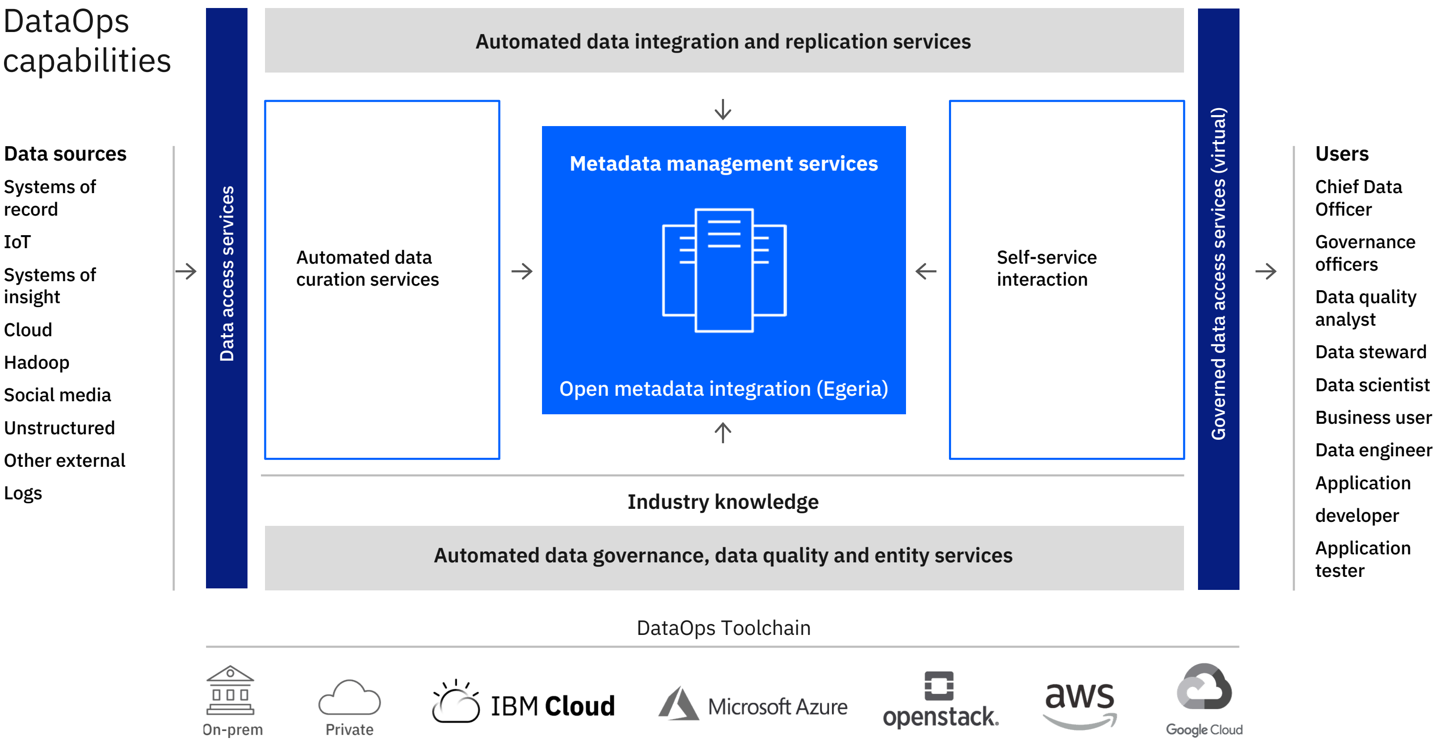


Figure 2: Examining an information architecture in support of DataOps

The provision of business-ready data includes all of these aspects, and any DataOps practice must include a holistic approach incorporating all 5 aspects. Organizations that focus on one element of the data pipeline at the expense of others are unlikely to realize the benefits from implementing DataOps practices. The technology conversation and implementation shouldn't be siloed from the ongoing planning regarding people and process. The tooling helps support and sustain the culture.

Bottom line

DataOps can seem daunting when organizations are still struggling with basic issues like defining data stewards' roles or creating data validation rules. However, the DataOps practice offers a solution to many failures that organizations have experienced in their digital transformation initiatives.

The most prevalent example of a failure that organizations recognize is within their data lakes. Many are on their second, third or fourth attempt to find technological success and have revitalized their leadership teams to take on the cultural changes required. *But why have these data lake implementations failed in the past?*

Many of these projects focused solely on ingesting uncleansed and ungoverned data into the data lake. Most likely, that failure has occurred due to limitations in effectively addressing people, process and technology issues effectively.

IBM DataOps program

The shift to adopt DataOps is real. According to a recent survey, 73% of companies plan to invest in DataOps.[2] IBM provides a path to a DataOps practice with a prescriptive methodology, leading technology and the IBM DataOps Center of Excellence (CoE). In the CoE, IBM experts work with organizations to customize an approach based on business goals and identify the right pilot projects to drive value for stakeholders.

IBM DataOps capabilities help deliver business-ready data by providing industry-leading technology that works together with AI-enabled automation, infused governance and a powerful knowledge catalog to operationalize continuous, high-quality data across the business. It serves to increase efficiency, data quality, findability and imbues governing rules to provide a self-service data pipeline to the right people at the right time from essentially any source.

From solutions that facilitate governing data lakes to developing applications and helping ensure regulatory compliance, IBM DataOps helps organizations show the value of data on optimizing decisions and time. Delivering high-quality enterprise data to enable AI is within reach—when an organization knows, trusts and uses data to drive value in the cloud and in any critical environment.

The IBM Cloud Garage methodology meets DataOps

The IBM Cloud Garage methodology is an approach to enable business, development and operations to continuously design, deliver and validate new functionality. The practices, architecture and toolchains cover the entire product lifecycle from inception through capturing and responding to customer feedback and market changes. The Open Toolchain architecture is designed to make it easy to combine IBM Cloud™ Platform services, such as Continuous Delivery (CD), with open source and leading third-party tools into an integrated toolchain aligned with DataOps practices. These patterns can be shared between teams as templates to promote successful adoption of DataOps across an organization.

IBM has identified six phases in the DataOps lifecycle, plus necessary cultural considerations, for successful implementation of a DataOps practice. This is based as well as internal DataOps adoption as part of the transformation journey for IBM.

The IBM Cloud Garage methodology describes these six phases as:

- **Think.** Conceptualization, refinement and prioritization of capabilities
- **Code.** Generation, enhancement, optimization and testing of features
- **Deliver.** Automated production and delivery of offerings
- **Run.** Services, options and capabilities required to run
- **Manage.** Ongoing monitoring, support and recovery of offerings
- **Learn.** Continuous learning and feedback based on outcomes from experiments

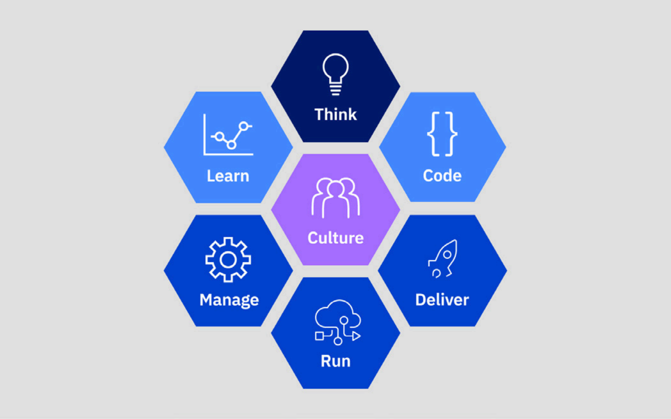


Figure 3. The six phases of IBM Cloud Garage methodology

Think: Continuously assess your DataOps maturity and align it to business goals

DataOps can be transformational to an existing organization and established processes. The intention of DataOps is to automate many existing manual tasks and streamline the data pipeline creation process. Whether starting or sustaining a basic DataOps practice, it is important to assess the teams' ability to deliver business-ready data fast and make a plan for improvement that aligns with creating business value.

DataOps success begins with cataloging data assets by capturing metadata and assigning policies to data classes, assessing and scoring data quality, and leveraging tools for integrating data as opposed to spreadsheets, tribal knowledge, or hand coding. When the team's maturity level has been defined, the goals and objectives should be to improve capabilities across as many DataOps aspects as possible.

DataOps teams should focus on aligning the delivery of necessary data with the value it can bring to the business. Ask the question: *How much money could be saved or made if this information were made available quicker?*

Code: Use a version control system—source control management

A data pipeline is source code that's responsible for converting raw content into useful information. This pipeline is core to data analytics and can be automated end to end to produce a source code that can be consumed in reproducible form. Different files, configurations and parameters associated with analytics are distributed in various places and environments within an organization without any governing control, which leads to inconsistent deployment. A revision control tool, such as GitHub, helps to store and manage all of the changes to code and configuration. A centralized repository also helps enterprises to have consistent and reliable information every time across environments, including any eventuality or disaster with a reliable recovery. Revision control also helps teams parallelize their development efforts and enable them to be agile in their delivery pipeline by using branch and merge.

To be certain that the data analytics pipeline is functioning properly, it must be tested. With continuous integration/continuous development (CI/CD), complemented with parametrization, it can be deployed and tested in a completely automated manner. Testing of inputs, outputs and business logic must be applied at each stage of the data analytics pipeline and checked for its accuracy or potential deviation—along with errors or warnings before they are released ensuring consistent quality. Manual testing has no scope in high-performance organizations, as it's error-prone, time-consuming and laborious. A robust, automated test suite is a key element in achieving CI/CD and essential in the on-demand economy.

Deliver: DataOps process and workflow automation—data technologies

For a DataOps methodology to be successful, automation is essential and requires a data analytics pipeline designed with run-time flexibility. A key requirement for the delivery of trusted data is a governed and consistent data pipeline that relies upon data curation, data ingestion, catalogs and classification using metadata and data sampling techniques.

A repeatable and robust data pipeline for the delivery of trusted and governed data needs a mechanism to do the following activities:

- Define and enforce data governance and data privacy policies consistently.
- Support efficient data movement.
- Initiate remediation or have industry-specific best practices and templates with predefined glossaries.

This process can deploy a governed data pipeline across different platforms consistently without any source code or configuration changes and deliver fully governed and trusted data. The DataOps process also needs to be supported with appropriate tooling for remediation to support exception handling and management. Back tracing and auditability of any change is a defacto requirement for these governed data pipelines.

IBM offers new innovative capabilities that include embedded machine learning (ML), AI-automation, infused governance and a powerful data catalog to operational continuous high-quality data across the business. DataOps efficiency depends on extreme automation of data technology components used for the data pipeline.

- IBM Cloud Pak for Data®, including IBM Watson® Knowledge Catalog (WKC) can address these requirements in an efficient, robust, automated and repeatable manner.
- IBM Cloud Pak for Data Server can address the need for data movement, publish and use within a data pipeline while helping to ensure data quality and policy enforcement. With efficient source control management, it can be automated and executed efficiently from within the CI/CD pipeline.
- Built-in ML within IBM Watson Knowledge Catalog for IBM Cloud Pak for Data complements the automation process and optimizes it with every iteration for robust pipeline.
- IBM Cloud™ DevOps Insights can help in providing operational insight and visualization for the data pipeline. It helps enforce security and quality measures that are monitored continuously, detect any unexpected variation and generate operational statistics based on extreme automation and customized integration with IBM Cloud Pak for Data.
- Apache Airflow and NiFi can help with workflow design and its orchestration.
- Use of extreme automation using REST endpoints along with parametrization, can help in selecting specific data sets or an environment dynamically and alter the behavior without impacting the source code of the pipeline and accommodate the day-to-day needs of data analytics professionals.

Run: Continuous integration and deployment

Continuous integration
Data pipeline engineers or owners can make updates or changes to the pipeline anytime and keep it within the revision control system as a private copy within the development branch or private branch. Multiple engineers can work in parallel and deliver changes to development or private branches concurrently and boost the productivity multifold. When pipeline changes are completed and tested within the branch, the source code can be merged into the main code base or trunk and delivered into the production line. In the event the merged code doesn't work, the data pipeline can always fall back into the previous working version of the pipeline source

code. Branching and merging allow the data analytics team to run their own tests, make changes, take risks and experiment and discard if a set of changes proves to be unsuccessful.

Continuous deployment

Data analytics professionals require relevant data used by pipelines separate from the private copy of source code and an environment to execute these pipelines. Working directly on the production database or environment isn't efficient and often leads to conflicts. To reduce conflicts and dependencies, data pipelines require:

- Efficient source control management
- Flexible environment deployment option availability
- Testing data behavior

Jenkins Pipeline is a tool that complements the IBM delivery pipeline. Red Hat® OpenShift® provides a repeatable, consistent deployment platform to be used for validating concurrent instances of a data pipeline with different values provided at runtime.

Manage: Work with consistent frequent deployment

Data analytics professionals want to avoid deploying changes that break the current data pipeline in production. Two key workflows can address the situation:

- **Value pipeline.** Creating value for organizations now with data flows into production.
- **Innovation pipeline.** Leading to the future of new analytics that undergo development and is added to the production pipeline.

Both these pipelines intersect in production where DataOps organizations master the orchestration of data to production and the deployment of new features while maintaining impeccable quality. Data pipeline quality control, such as statistical process control monitoring the data and new development pipelines, the development team can deploy without worrying about breaking the production systems. With agile development and DevOps, the velocity of new analytics is maximized and fast. It helps to minimize the time and effort to turn a business need into an analytic idea and release it as a repeatable and reusable production process.

Learn: Communication and process management

Efficient and automated notification is core to the communication and remediation process within a DataOps methodology. When changes are made to any source code or when the pipeline is triggered, fails, completed or deployed, it can be notified. In the event of failure, information can be pushed along with notification to address the problem. The post remediation process can be triggered automatically to validate the pipeline, deploy it into the next stage and update the dashboard with the latest information and data quality. Tools like Slack, Apache Kafka, PagerDuty and Trello are commonly used to facilitates cross-stakeholder communication, collaboration, feedback capture and sharing as part of the DataOps toolchain.

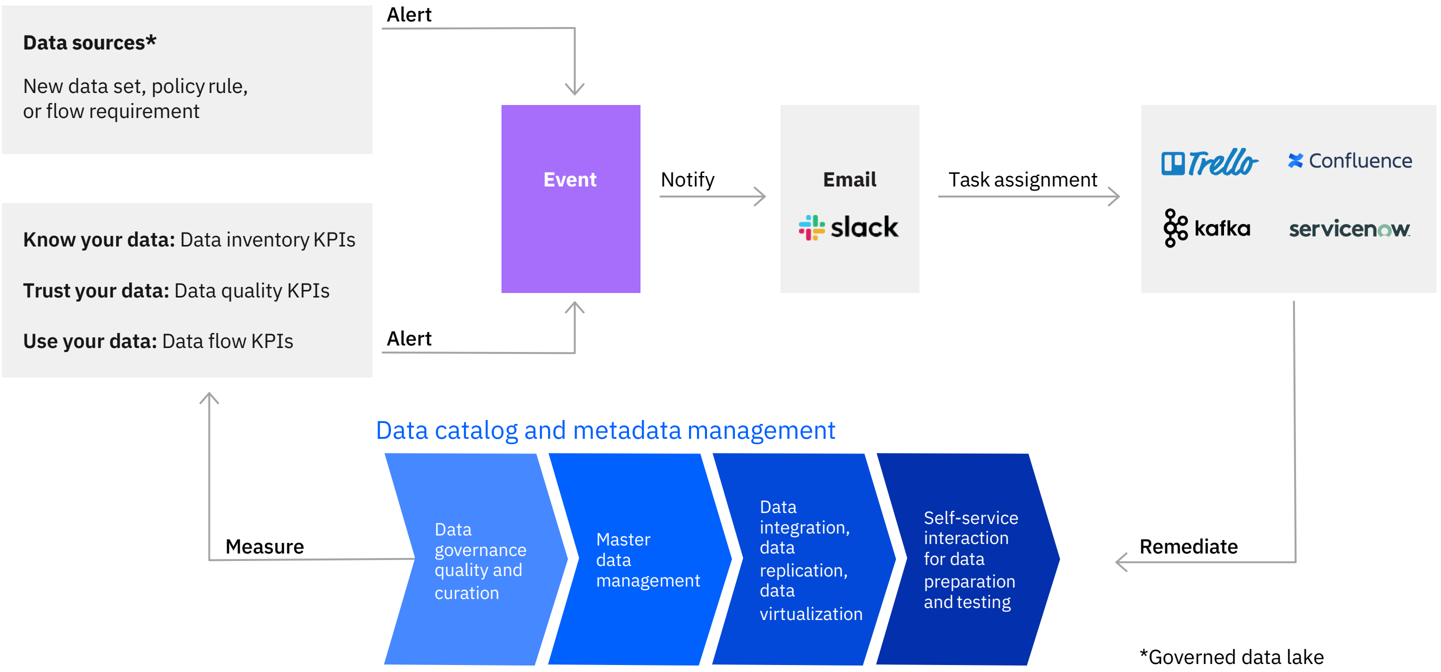

Data sources*
New data set, policy rule, or flow requirement
Alert
Know your data: Data inventory KPIs
Trust your data: Data quality KPIs
Use your data: Data flow KPIs
Alert
Event
Notify
Email
slack
Task assignment
Trello
Confluence
kafka
servicenow
Remediate
Data catalog and metadata management
Data governance quality and curation
Master data management
Data integration, data replication, data virtualization
Self-service interaction for data preparation and testing
Measure
*Governed data lake

Figure 4: Visualizing communication and process management within a governed data lake environment

The impact of a successful DataOps practice

By adopting a DataOps practice, a retailer made improvements across its data pipelines so that data changes took less than two minutes to be applied across the organization versus previously taking three weeks. As a result, the retailer leveraged business-ready data to conduct customer affinity analysis in less than one day—a process which previously took 20 days. Furthermore, it reduced the amount of time needed to report on inventory stock positions to one sixth of the time.

The criteria for a successful DataOps practice include:

1. **Establishing a data office.** This process involves firmly defining the scope of the role in providing data as a resource to the bank, identifying the key executive stakeholders, and understanding the commitments that every stakeholder in the data pipeline makes to a collaborative operation and culture.
2. **Aligning with business objectives.** To remain competitive, the market necessitates rapid response to new opportunities that can only be delivered by an informed and data-led approach. In short, unless there's excellent communication between business and data delivery, leadership knows that their organizations will not thrive.
3. **Scaling data successes.** With every data-led initiative, leaders need to ensure that the data produced can be used and reused continuously, with value increasing each time. This result can only be achieved when it's shared centrally, is searchable and aligned with a business language.

Conclusion

Organizations that have successfully deployed DataOps know what data assets they have access to, trust the data meaning and its quality, and use their data to its maximum potential. Data has value when trusted business-ready data helps drive differentiated insights, operational excellence, collaboration and competitive advantage.

Establishing a DataOps practice requires:

– *Investing in understanding their organization's unique competence and challenges by running a pilot project*

– *Using the pilot project's success to expand and grow DataOps skills and the organization*

– *Promoting its success to recruit more teams to participate in the DataOps practice*

– *Sharing lessons learned and start to build a DataOps CoE*

Take the next step and schedule your own IBM DataOps Garage Workshop and accelerate your path to business-ready data by contacting dataops@us.ibm.com.

Organizations are finding they're well-positioned to adopt DataOps if they have already been working on a data catalog, data lake or master data initiatives. Learn more about DataOps support with market-leading technology at ibm.com/DataOps.

IBM Corporation
New Orchard Road, Armonk, NY 10504

Produced in the United States of America
April 2020

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed, misappropriated or misused or can result in damage to or misuse of your systems, including for use in attacks on others. No IT system or product should be considered completely secure and no single product, service or security measure can be completely effective in preventing improper use or access. IBM systems, products and services are designed to be part of a lawful, comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM DOES NOT WARRANT THAT ANY SYSTEMS, PRODUCTS OR SERVICES ARE IMMUNE FROM, OR WILL MAKE YOUR ENTERPRISE IMMUNE FROM, THE MALICIOUS OR ILLEGAL CONDUCT OF ANY PARTY.

[1] 2019 Global data management research: Taking control in the digital age." Experian, 2019.

[2] Jarah Euston, "The DataOps Trend is Real: 73% of Companies Plan to Invest in DataOps to Manage Data Teams in 2018," Nexla

10028810USEN-01

Appendix: DataOps Pilot Program Template

Project name:

Date:

Department or unit:

Pilot program leads:

Name	Role	Email	Phone

Extended stakeholders across the business:

Name	Role	Email	Phone

Problem statement:

Root cause checklist:

Challenge	Applicable? Y/N	Additional notes

Metrics for success:

Start date: Sprint end date:

Assessment:

Implementation	Current state	Desired sprint outcome	Action steps to achieve the desired outcome
Data asset ingestion, automated discovery and classification			
Data quality assessment and remediation			
Business terms assignment			
Data privacy, regulatory compliance and corporate policy definition and enforcement			
Data consumer requirements definition request processing			
Data request communication and notification, including exception and error handling and remediation			
Curated data publication into the catalog			
Data lineage and reporting			
Collaboration, feedback and audit			

Example questions to ask during an implementation audit:

Data asset ingestion, automated discovery and classification

- Do we perform high-volume, low-latency replication for business continuity?
- Do we use advanced streaming analytics for real-time, low latency analytics?
- Do we easily connect to any data source and perform complex data transformations and integrations?
- Can we provide data from social media, weather data or other public cloud data sources?
- Do our data consumers have real-time access to our metadata repository from any desktop application?
- Do our data consumers have real-time access to our data catalog to allow self-service and assistance in the discovery of data sets that are relevant to their job?
- Do we use data profiling tools to understand the data, validate data values, column and table relationships and find and analyze anomalies?
- Is our business rules management integrated with our metadata management infrastructure?