HITACHI
Inspire the Next

WHITE PAPER

Is DataOps Your Windfall of Value?

By Hitachi Vantara

August 2019

Contents

Is DataOps Your Windfall of Value?

The Opposite of DataOps

The current state of data management is riddled with problems and bottlenecks that block data from delivering full value. The growth of cloud, mobility, expanding governance requirements, new data sources from IoT and the edge and increasing use of AI and machine learning (ML) have all changed the game. There is more data and more ways to use it than ever.

To support the goal of monetizing data, the modern data supply chain must enable analysis of new types and combinations of data at a far greater scale than has been possible with traditional enterprise business intelligence and data warehousing systems.

Use cases like AI/ML need to create a data pump that brings data from sources, transforms it to the desired form and keeps it flowing. But the same is true for almost every other use case. In fact, one can see the modern data supply chain as a huge number of data pumps, often called data pipelines, providing just what is needed for analytics, applications, automation, governance and optimizing processes of all types.

To address this challenge, companies must reimagine data management practices for the modern era. At Hitachi Vantara, we believe that the most comprehensive and coherent vision is an approach called DataOps.

DataOps streamlines and automates the creation, governance and operations of a vast data supply chain, repeating the success of DevOps. When properly implemented, DataOps seamlessly connects data consumers with data experts to increase communication, collaboration, governance and accountability. The result is a unified process and platform that delivers timely and accurate business decisions and operational efficiencies.

What we have now is in many ways the opposite of DataOps, a system rife with barriers and separations, with priorities set without regard to the needs of users. While highly skilled experts maintain the current data supply chain, complexity has gotten out of control.

Most current data supply chains require skilled staff who focus on issues such as data quality, data engineering, data governance and data profiling. Experts use a number of disjointed tools that are both off-the-shelf and developed in-house. This hodgepodge of technology and processes is used to take data from many sources, shape it and move it to where it is needed.

The resulting data management infrastructure is complex, expensive and brittle. The process of locating and blending data to solve a problem is slow and highly intermediated. Though all parties are doing their best, communication, collaboration and self-service are lacking.

These difficulties prevent people from finding the data needed to address pressing business issues and it is nearly impossible to keep pace. Perhaps worst of all, those who apply the data practically in the business cannot collaborate closely with data experts, who have detailed knowledge of what data is available and how to put it to use, but are often overwhelmed with requests.

To answer these challenges, DataOps delivers the right data to the right place at the right time. Doing so consistently at scale for all data means crossing the vast chasm between those who require quality data to gain business value and technical experts who transform raw data into a useful form.

We must redesign and reimplement not only the flow of data, but how the flow is governed and how supporting operations infrastructure works. This requires increased automation, expanded use of metadata and policy-based approaches to reducing complexity.

The companies who achieve even partial success in moving toward a DataOps vision will gain a windfall of value from increased use of data and the cultural changes that flow from such a transformation.

The Mission of DataOps in Three Layers

Before delving into how DataOps can transform businesses, it is helpful to define it and look at its origins. DataOps is enterprise data management for the AI era, enabling you to connect data consumers and creators to find and use the value in all your data.

DataOps is not a product, service or solution. It's an approach, a technological and cultural change to improve the use of data through better data quality, shorter cycle times and streamlined data management.

DataOps builds on the success of the DevOps movement, which obliterated the divide between software development and operations and as a result increased the quality of code, expanded automation of the development-to-operations cycle, shortened delivery times and increased the pace and responsiveness of innovation.

DevOps radically altered the process of software development, deployment and operations. Before DevOps, developers would create new software and applications and then pass them off to operations staff, who were tasked with keeping the software running. The DevOps process combined development and operations to create integrated teams of developers and operational staff. Working together, with the help of an automated, end-to-end tool chain, they would systematically create software in an integrated and iterative way focused on the ultimate goal: customer happiness.

DataOps applies the principles behind DevOps to the world of data management to create three layers that constitute a new data management infrastructure based on:

- Agile Data Pipelines
- Governance and Instrumentation
- Operations Agility

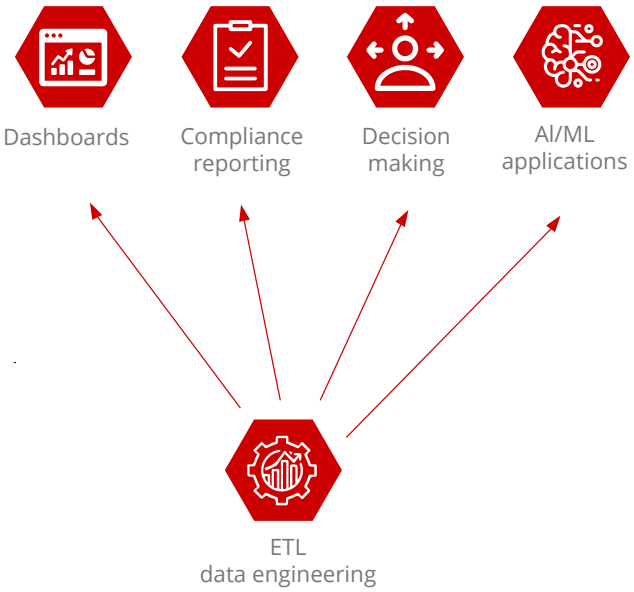
Layer One: Agile Data Pipelines

DataOps dismantles the barriers that frequently exist between data experts and business users who need data to serve customers. Instead of having data experts push data toward end users, in DataOps the end users who want to solve a business problem gather the data needed, "pulling" it from wherever it is located. The entire data supply chain is reshaped to support this self-service process for finding data, blending and reshaping it and creating a data pipeline.

It is easy for anyone experienced in the history of business intelligence to get the wrong idea about what we mean by agile data pipelines. The development of data dashboarding and discovery analytics environments, simplified ETL and data catalogs created a situation where an analyst could find and assemble data and in some cases have that data delivered regularly. But creating a data pipeline for analytics is just one type of data pipeline. We will need different data pipelines for AI/ML models, applications, chatbots, support for automation, compliance reporting and so on. The end-users who need this data must be able to create these data pipelines, put them into production and adjust them as needed. Self-service is not just about analytics but about creating a broad set of agile data pipelines.

Agile Data Pipelines

Users pull the data they want for:



Goals

- More people can create data pipelines
- Self-service creation of data pipelines
- AI/ML assists in self-service process
- Guided search of data catalog
- Simplified operations
- AI/ML model management

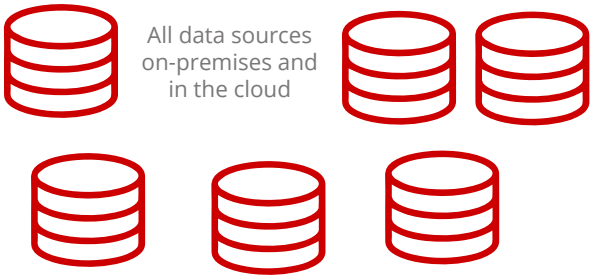
Governance and Instrumentation

External and internal compliance can be automated and security applied

- Fine-grained access control
- Sensitive data masked
- Operational monitoring
- Automation of compliance
- Policy-driven data security
- Usage monitoring

Operational Agility

Policy-based configuration increases operational agility



- Decreased cost and complexity
- Control data management using policies
- Expanded automation
- Increased agility

For this to work, DataOps infrastructure must be far more searchable and automated than current data management infrastructure. Users must be able to easily to locate data, transform it into the shape needed for a use case and create a data pipeline to keep it flowing. Because the amount of data and the demand for data pipelines is growing rapidly, the number of people who can do this work must also grow rapidly. With DataOps, self-service is how this happens, replacing the push model of the current data supply chain with a pull model.

PUSH MODEL		
Data experts and data engineers create and optimize pipelines to deliver data to the business.	The technology of data management is used to create and optimize data pipelines.	End users make use of the data provided to them, making requests to improve the data infrastructure as needed.
PULL MODEL		
Data experts and data engineers support self-service for finding and transforming data to create data pipelines.	The technology of data management is used to increase the amount of data in the platform and make it easier to find, transform and use data.	End users advance the use of data by working on their own and collaborating with data experts.

The operational characteristics of data pipelines vary dramatically. Some will be slow-moving, time-insensitive batch processes that need to be run once a week or once a month. Other pipelines may need to be in real time, making use of streaming technology. The configuration of data pipelines must allow all types of data pipelines to be created and maintained.

The ultimate goal is to make data more useful and derive more value from it. To do this, data must be stored, enriched and governed in new ways that support the DataOps process.

Layer Two: Governance and Instrumentation

It is vital to recognize that the challenge of DataOps is more complex than the world of DevOps, which improved the automation between operations and development in a far simpler landscape.

DevOps was an engineer-to-engineer process in which developers and operations people learned to work better together. DataOps will involve data experts, IT, operations, data analysts, data scientists, end users – literally the entire company and possibly even its customers and business partners.

In the world of DevOps, the assets in play were the code used to create software, the configuration of the operational environment and the configuration of the toolchain used to automate testing and deployment.

In DataOps, there is a larger landscape of data coming from many different sources with different characteristics. Allowing universal access to all this data is not an option. Use of data with personally identifiable information is usually regulated and its usage must be tracked. Other types of data may be just as sensitive or confidential. For these and many other reasons, DataOps requires a highly sophisticated form of governance with techniques for controlling access, masking and anonymizing data and other capabilities.

In addition, the entire DataOps infrastructure must be instrumented for several reasons: First, reporting on access and usage may be required to comply with regulations. Second, monitoring of the agile pipelines is required to debug operational problems. Third, usage data also can be tremendously valuable in optimizing data pipelines, expanding exposure for highly used data and directing investment in acquiring new data. It is likely that metrics used in manufacturing such as cycle time, quality measures, mean time between failure (MTBF) and mean time to repair (MTTR) can be used to measure the data supply chain.

Layer Three: Operations Agility

Pipelines are the way that data moves from repositories to environments where it becomes useful. Operations supports both creation and implementation of pipelines but also the general management of all data.

There are a wide variety of repositories and storage mechanisms that reside in the modern edge-to-core-to-multicloud infrastructure. Data is accessed and stored in IoT devices, branch offices, edge computing, on-premise data centers, co-location facilities, a variety of SaaS applications and public cloud infrastructure and other distributed environments as well. Just as the creation and operation of data pipelines must become more simplified and abstract, so must the underlying data storage and management infrastructure.

The abstraction and simplification required for assembling data and creating and modifying data pipelines must extend to data storage and management infrastructure. The only way to have visibility and control across the distributed continuum of data repositories is to adopt a metadata-driven strategy. The metadata must describe the contents of each data set, where it is stored and its risk profile. It must also hold the details of the required data management policies such as retention, storage optimization, SLAs for performance, encryption, security and governance. This metadata can drive automation of the DataOps infrastructure so that properly tagged data is managed and delivered in the correct way. For example, data with high performance SLAs may need to be stored in tiered storage with high performance in-memory buffers. Data marked for retention must be preserved even if logically deleted. Data that is rarely accessed may be archived to reduce costs.

For all of these reasons, the DataOps infrastructure, like DevOps, will have to break new ground in terms of the scope of automation and the flexibility of configuration. When a data infrastructure reacts to metadata in this way, sophisticated management can take place through metadata-driven policies and configuration without hugely complex configuration and code that needs to be maintained. Getting to such a DataOps infrastructure is not easy, but once there, a company gains tremendous power.

Hitachi Vantara's vision for DataOps consists of an infrastructure that implements these three layers. But the success or failure of DataOps rests heavily on how these layers are implemented. If coding by highly skilled experts is the only way to create and maintain the data supply chain, DataOps will never work.

How DataOps Will Change Technology

For DataOps to succeed, the implementation must not create vastly complex data pipelines and data management infrastructure that can only be curated and improved by experts. Instead, like it has in DevOps, the scope of automation must expand from the beginning to the end of the process with automation and configuration that takes place through policies rather than coding.

At Hitachi Vantara, we see complexity being conquered by expansion of metadata, automation and policy-driven control and configuration.

Expansion of Metadata

In a DataOps platform, data is far more richly described with metadata to support better search, self-service data assembling and blending, configurable data management infrastructure and a variety of forms of automation. There can be no more dark data. By describing data with an expanding collection of metadata, we bring data into the light.

When attempting to locate data to solve a business problem, the expanded use of metadata in DataOps allows users to describe the business problem and determine what data is available to help. The metadata allows advanced search systems to locate relevant data. As a DataOps platform matures, formal semantics models can be added to make search and infrastructure even more powerful.

Metadata also supports expansion and automation of governance processes. By identifying which data is sensitive, policies and data collection can be implemented to control and monitor its use.

Better metadata is also the foundation for implementing policy-based automation and configuration of data management infrastructure. For example, if you tag certain data to be stored in a particular geography, the system can automatically make sure that happens, instead of implementing (and updating) such rules in complex code.

Expansion of Automation

Expansion of metadata is table stakes for DataOps. The expansion of automation and policy-driven control and configuration are how DataOps breaks new ground.

In DevOps, automation meant that you could deploy an entire complex website or application, starting from source code through compilation, testing and deployment across a complex operational infrastructure by pressing one button. This was only possible because each stage of the process had been automated. In the past, such work was attended to by experts at every step.

In DataOps, automation will work the same way. Tasks that are complex now will take place by pressing one button. Sequences of complex tasks will be connected in larger automatic pipelines.

In DevOps, this transition took years and required the development of new technology such as continuous integration/continuous deployment (CI/CD) systems. The way software was developed changed as microservices and containers became more widely adopted, making DevOps processes easier to implement.

DataOps is at the beginning of this transition, but we can already see progress in how the systems for ETL, data prep, data quality, controlling data pipelines, repositories and data management infrastructure are becoming more automated and more self-service.

Expansion of Policy-Driven Control and Configuration

There is a lot of complexity in the current data landscape that just won't go away. But that doesn't mean that the complexity cannot be conquered. It will be through policy-driven control and configuration that describes to the system what work should be done.

For example, today's automobiles are incredibly complex and have many subsystems that deliver advanced services for automatic braking and navigation, but the cars are manageable and easy to operate. Tomorrow's data management infrastructure will be the same way.

The steering wheel and dashboard for a data management infrastructure will be a set of policies for how a system should behave or for how to set high level configuration rules. These policies can only work after a domain is well understood, has a full set of API-controlled services that do the important work and has automated many complex tasks. Work is carried out as these services are configured and orchestrated.

At that point, it is possible to set a policy in metadata, such as "store this table in Germany," and have it carried out.

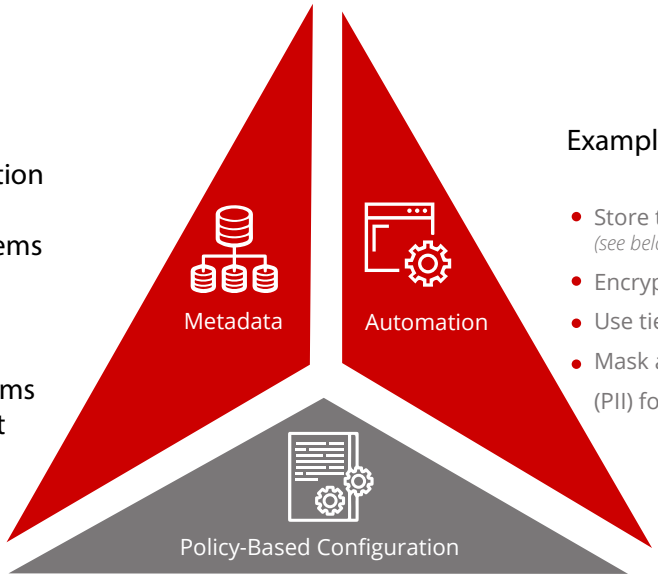
This is what happened in the world of DevOps as services for automating each step of the processed emerged and were then orchestrated into an end-to-end process.

In DataOps the landscape of services is far deeper and wider, but it is already clear for selected use cases and domains that such policy-driven control and configuration are possible. This transition will take time, but DataOps will thrive as the scope of automation that can be controlled and orchestrated by high level policies expands.

How Policy-Based Configuration Reduces Complexity

Policy-based directives empowered by automation and metadata simplify control of complex systems

Policy-based directives allow us to control systems by stating what we want the system to do

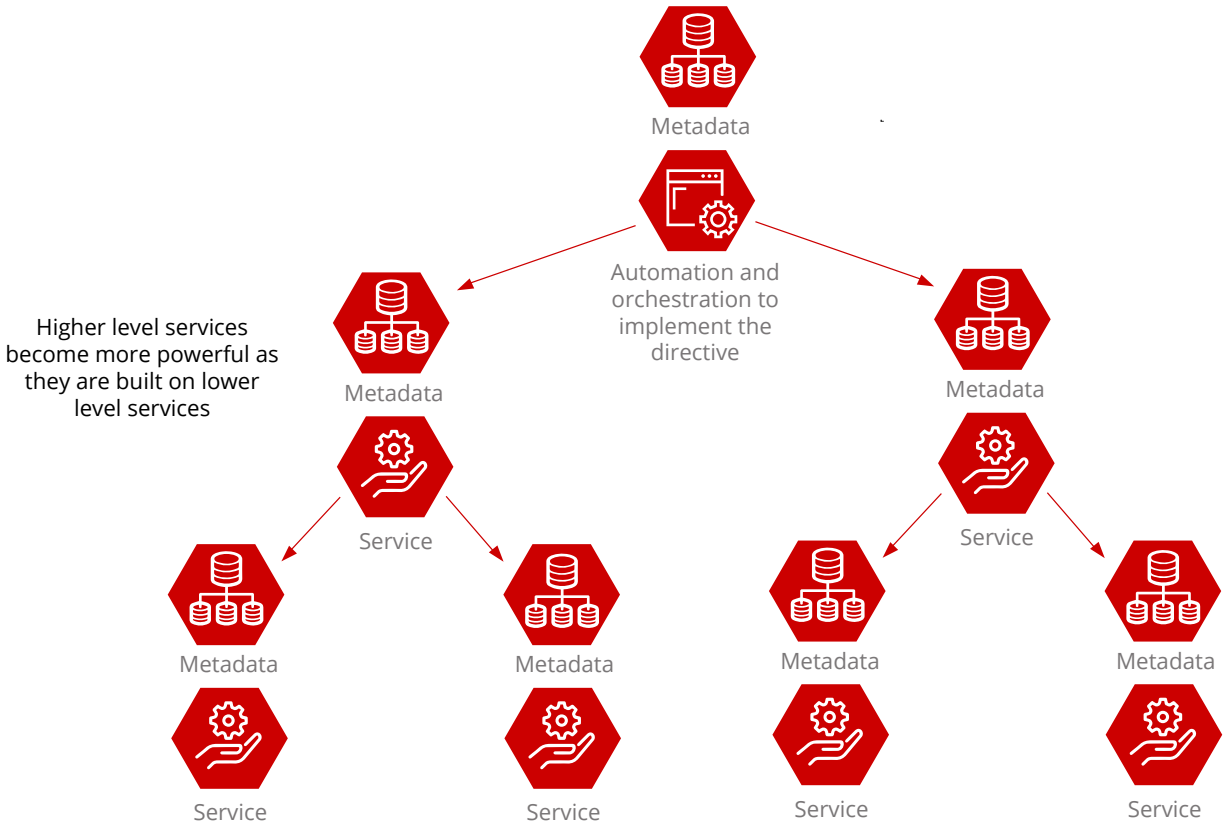


Examples

- Store this database table in Germany *(see below)*
- Encrypt this database
- Use tiered storage for this file
- Mask all personally identifiable information (PII) for this group of users

Example: Store this table in Germany

Policy-based directives instruct automated systems to carry out the policy using a collection of services that rely on metadata



The result? Complex work can be described simply, increasing agility.

Key Observations

The DataOps vision is compelling and will drive companies using data and vendors creating products for years. It is important to recognize important aspects of the journey.

- Existing technology and skills will be part of DataOps. DevOps was not a “big bang” refresh, but an evolution. DataOps will be the same way. The practices of data engineering, data quality, data governance, master data management and data catalogs will all be part of the world of DataOps and will be deployed in new ways to support the DataOps vision.
- Just as CI/CD systems developed to become the orchestrating brain of DevOps, similar new technology will bring DataOps to life and support processes such as AI/ML model development, deployment, monitoring and management that are becoming increasingly important.
- DataOps will fail if a focus on technology obscures the goal, breaking down barriers between data consumers and data experts. DevOps succeeded not just because of an automated toolchain but because developers and operations staff started to collaborate and understand each other. In DataOps, the way the technology is built and deployed must be pulled by business needs, just as in DevOps the shape of the product is pulled by signals about customer happiness.

DataOps is a bold vision but one that is worth fighting for. At Hitachi Vantara, we are dedicated to accelerating the development of people, processes and technology to achieve this vision. We are eager to help companies who see value in DataOps and want it to work inside their organizations.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
HitachiVantara.com | community.HitachiVantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
HitachiVantara.com/contact