2024 Planning Guide for IT Operations and Cloud Management

Published 4 October 2023 - ID G00800386 - 55 min read

By Analyst(s): Gregory Murray, Ang Troy, Ross Fomerand, Andre Bridges, Dylan Roberts, Paul Delory, Alexandru Giurovici, Lydia Leong, Douglas Toombs, Steve White, Marco Meinardi, DB Cummings, Stanton Cole, Fintan Quinn

Initiatives: IT Operations and Cloud Management for Technical Professionals

As IT teams adapt to increasing levels of innovation and complexity, they also have to adjust to traditional roles and tools changing for greater collaboration and alignment with consumers and stakeholders. I&O technical professionals must prepare for these shifts and capture beneficial change.

Overview

Key Findings

- Although generative artificial intelligence (GenAI) hype and interest have exploded, adoption of the technology introduces some unfamiliar challenges for I&O technical professionals. Developing durable and scalable practices requires an early adopter mindset, focused internal investigation and iterative experimentation.
- The pressure to enable frictionless self-service for IT consumers is driving many innovations and practices to make it easier to establish internal platforms.
- The shift in operations that started with DevOps is expanding and accelerating. Collaboration and accountability are becoming increasingly focused on business metrics that require cross-functional expertise.
- Tool vendors are capitalizing on this shift to grow their addressable market by tackling the needs of cross-functional and adjacent roles.
- As the number of external services and data volume grows, layered resilience practices are an effective augmentation of good architecture and development hygiene to protect against malicious actors and failures beyond your control.

Recommendations

For I&O technical professionals:

- Start experimenting with GenAI, if you haven't already. Use it to prototype automation code as a way to accelerate and expand your automation skills. Propose GenAI experiments to assess opportunities and the potential to augment service delivery and meet internal customer requirements.
- Adopt platform engineering to maximize the flow of value from self-service teams, and capture the potential of managing IT delivery using proven product management practices.
- Realize the full potential of cloud adoption with specialized operations teams, processes and skills. Don't try to replicate cloud operations with on-premises, and don't use on-premises approaches to run your cloud environment.
- Embrace the shifts in operational accountability. Help cloud consumers understand and manage their consumption. Collaborate across functions to augment security operations and service management.
- Elevate resilience in your automation and system designs. Take advantage of the unique attributes and architectures of cloud-native applications to ensure availability, even when failures are malicious or beyond your control.

Strategic Planning Assumptions

By 2027, 90% of enterprises will use some AI functions to automate Day 2 network operations, compared with fewer than 10% in 2023.

By 2026, generative AI technology will account for 20% of initial network configuration, up from near zero in 2023.

By 2027, more than 70% of enterprises will adopt a centralized platform engineering and operations approach to facilitate DevOps self-service and scaling, which is a significant increase from less than 25% in 2021.

By 2027, platform engineering principles will influence more than 50% of infrastructure and operations technology decisions, which is a substantial increase from less than 20% today.

IT Operations and Cloud Management Trends

Tough economic conditions and demands for efficiency have driven some innovations in IT operations and cloud management in 2023. The resulting tension between multidisciplined efficiency and demands for autonomy continues to propel the cycle of continuous change in I&O. This means new ways of collaborating, new tools to enable intersecting roles, and the next wave of business requirements to capitalize on these shifts.

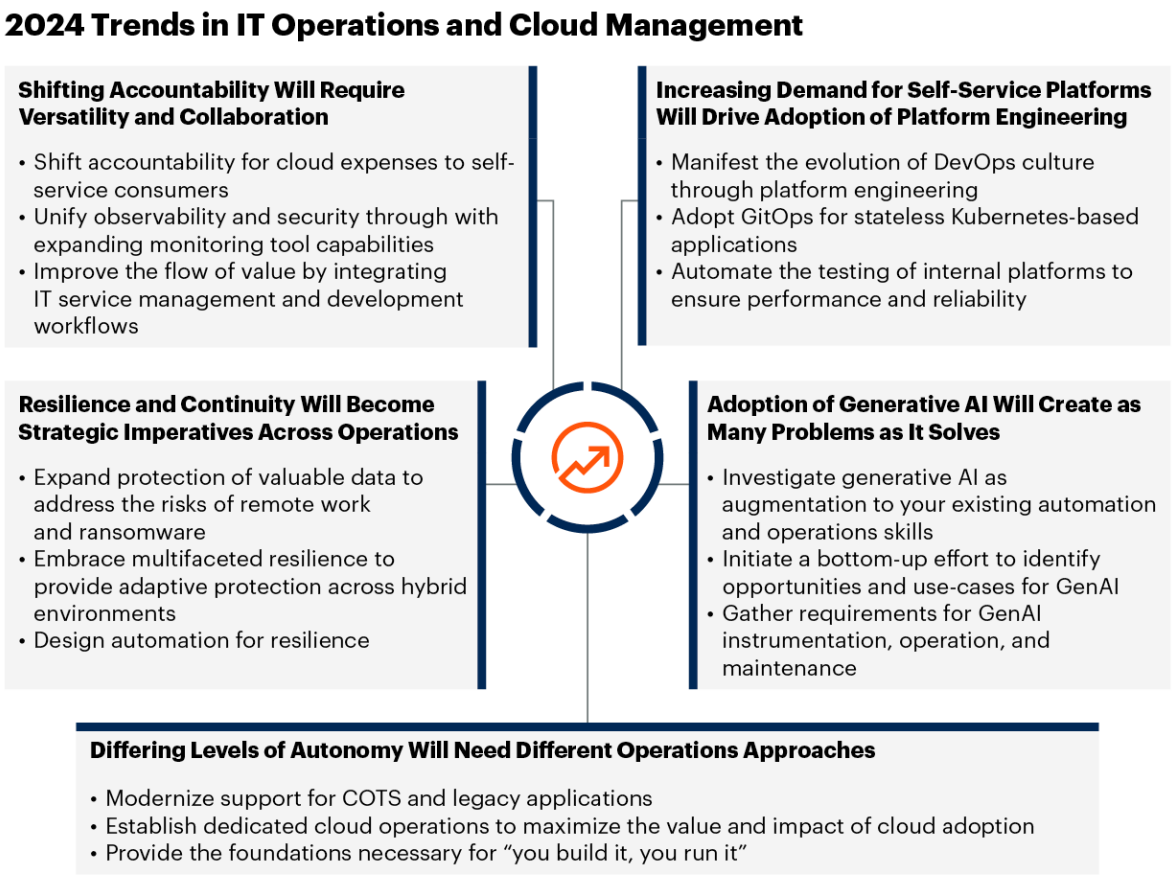
Without a doubt, one of the most hyped and interesting IT trends is generative AI and, in particular, large language models (LLMs — see Note 1). The level of interest and potential is unprecedented. Enterprises are feverishly trying to identify how to apply GenAI before their competitors do. If you're not already seeing it, expect to introduce GenAI within the next year. Even if you have no plans for adoption or have plans to avoid the risks and delay adoption, vendors in every market are working quickly to add GenAI features to products you're already using.

There are many interesting applications of GenAI within I&O; for instance, code summarization to help understand how software works, generating documentation to formalize policies, and creating incident summaries. There's particular interest in the GenAI systems that generate code, most of which can also produce code for the more common approaches to IT automation. Our experiments have found that these systems are not a total solution, but they're far more effective than searching the internet for sample code.

Platform engineering continues to be one of the Gartner top trends across IT. This approach brings the benefits of product development and efficiency to IT by focusing on incremental delivery of items in a prioritized requirements backlog. The past year has afforded both open-source and commercial developers time to mature their offerings, making platform engineering more accessible. In practical terms, this means greater adoption of Kubernetes (K8s) as the orchestrator of applications, proliferation of "as code" practices that foster declarative and idempotent automation, and demand for greater collaboration between test, software and I&O teams to ensure customer satisfaction.

However, as these internal platforms become more accessible and more common, we're seeing that not all teams are ready for the shift in accountability that comes with self-service autonomy. In these cases, I&O needs to redouble its efforts to ensure that operations approaches are kept up to date and, where necessary, tailored to the environment and consumers of those applications. In cases where self-service is taking hold, expect for a logical evolution that will push accountability as a way to augment autonomy. This should be a fairly natural transition and will result in more efficient, federated cost management, shared views of application security, and layers of resilience. Figure 1 captures these trends and the Gartner recommended actions and considerations for each.

Figure 1: 2024 Key Trends in IT Operations and Cloud Management



Source: Gartner
800386_C

Gartner

The 2024 technical trends that will impact most I&O technical professionals are:

- Adoption of generative AI will create as many problems as it solves.

- Increasing demand for self-service platforms will drive adoption of platform engineering.
- Differing levels of autonomy will need different operations approaches.
- Shifting accountability will require versatility and collaboration.
- Resilience and continuity will become strategic imperatives across operations.

Adoption of Generative AI Will Create as Many Problems as It Solves

Generative AI is a rapidly evolving set of AI techniques that learn from a massive set of data and generate original artifacts that reflect a likeness learned from the original data. Most technical professionals have already used a consumer GenAI application, such as ChatGPT, and are trying to understand the potential for augmenting and accelerating IT processes.

However, GenAI is neither a panacea nor a replacement for skilled, knowledgeable workers. While the potential improvements to productivity are considerable, I&O technical professionals must develop the skills to identify where GenAI is most impactful, how to engineer useful prompts, and how much trust they should put in the technology and its results.

Unskilled GenAI users may fall prey to blindly trusting results, misunderstanding inherent biases in data, or poor decision making because of hallucinations. Likewise, there may be personnel who have little to no trust in GenAI solutions, resulting in missed opportunities. Experienced users who understand the challenges and applications of GenAI will be linchpins in identifying and taking advantage of the best opportunities, and in disseminating best practices. [1]

More broadly, I&O technical professionals will be expected to support GenAI efforts throughout the enterprise. In addition to helping ensure compliance with regulations and policies governing its use, they will also be expected to manage the ongoing operation of GenAI and/or underlying components. Maintaining availability and the life cycle of complex GenAI solutions is neither simple nor a deeply developed market space. This will require engineers to implement observability, automate processes, allocate costs, track experiments, and manage ongoing problems and incidents.

To prepare for the adoption of generative AI, I&O technical professionals need to:

- Investigate GenAI as an augmentation to your existing automation and operations skills.
- Gather requirements for GenAI instrumentation, operation and maintenance.
- Initiate a bottom-up effort to identify opportunities and use cases for GenAI.

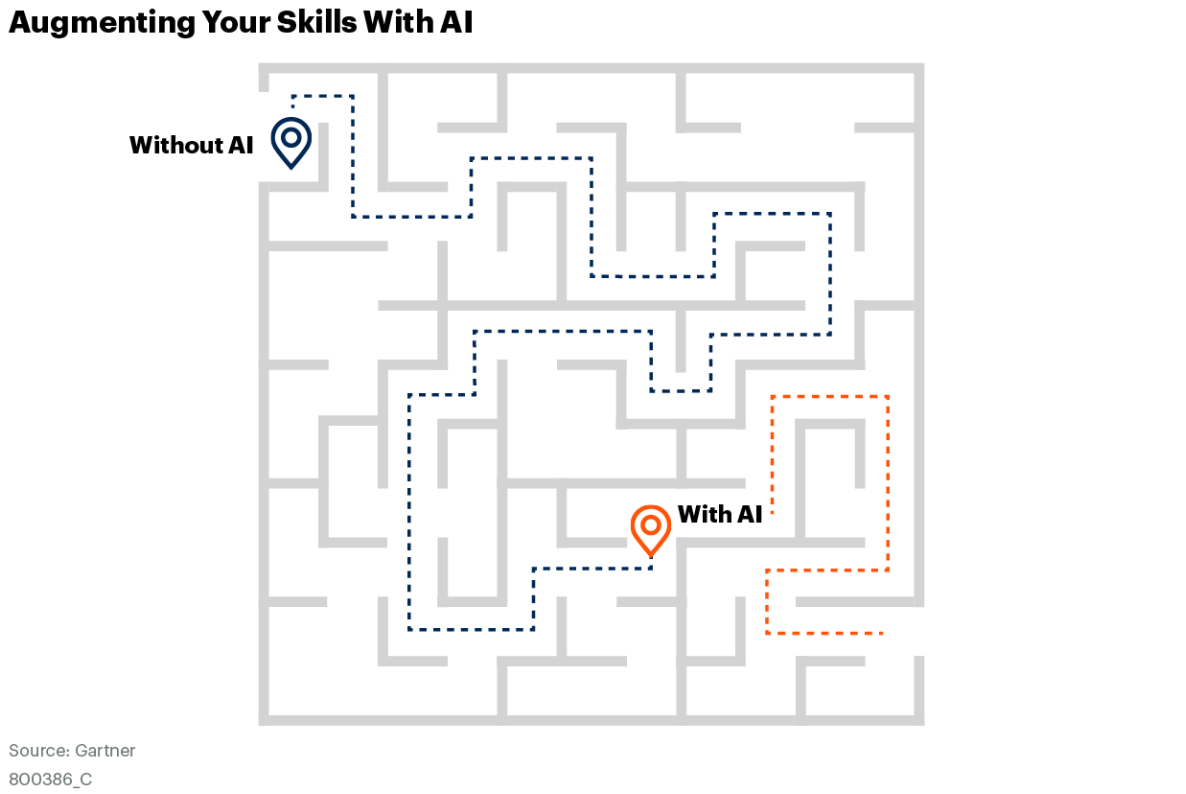
Planning Considerations

Investigate Generative AI as Augmentation to Your Existing Automation and Operations Skills

GenAI has afforded I&O professionals an unprecedented opportunity for the augmentation of current skill sets through services such as ChatGPT and Google Bard. Some of the most in-demand skills — automation, infrastructure as code (IaC) and continuous configuration management (CCM) — promise great opportunities for using GenAI to augment existing knowledge and accelerate learning.

You can develop these skills by using natural language queries to generate automation scripts, IaC manifests and CCM feature sets. Study these assets, use them as a starting point and learn to adapt them to your environment. As illustrated in Figure 2, think of these tools as helping solve the problem and getting you closer to the answer without having to work through the whole problem yourself. This will both augment your existing skills and help you discover more advanced automation, IaC and CCM uses.

Figure 2: Augmenting Your Skills With AI



Because GenAI is known to hallucinate and make errors, don't trust the output blindly. Instead, develop manual reviews and automated tests to ensure the generated code is tested and trustworthy in a fully staged continuous integration pipeline prior to committing to the master branch of the production Git repository. The quicker a junior engineer can implement creative solutions to complex problems using IaC and CCM, the faster a team's cognitive load will be reduced. Reducing your overall cognitive load will lead to greater overall productivity.

To take full advantage of these features, attend GenAI seminars, webinars and conferences, and deliver presentations based on what you learn. Most organizations' legal teams are concerned about intellectual property (IP) exposure or copyleft infringements from GenAI and are defining guard rails to protect IP, such as prohibiting the sharing of sensitive data or internal code with an LLM. Once a governance model is established, collaborate with security professionals to create an enforcement policy that allows for skill augmentation while simultaneously protecting IP.

Conduct an informal skill assessment to identify where you have and don't have skills to automate common activities. Use this assessment to create a roadmap of the target skills to augment and where GenAI can accelerate and develop your augmentation skills. Track your progress and development through personal goals and metrics. Some example KPIs would be:

- What percentage of your IaC or CCM was augmented by GenAI code generation?
- Using GenAI, how many hours writing IaC or CCM do you think you saved last month?
- What is the ROI or total cost of ownership of using GenAI to develop IaC or CCM assets?
- How many times did GenAI-generated IaC or CCM lead to an outage in production services last month?

Benefits

- Learning new skills is challenging and GenAI simplifies that challenge by making the learning more personalized to your scenario, your skill level and your individual tasks.
- Developing skills that use GenAI to create automations, IaC and CCM will make you more efficient in your existing role and will build valuable expertise in high-demand I&O skills.
- Augmenting existing automation skills with GenAI can amplify your impact, allowing you to solve more complex issues and tackle more use cases.

Cautions

- GenAI isn't a replacement for human skill and understanding, so don't view it as such. Instead, view GenAI in terms of augmenting human skill and understanding.
- The GenAI field is new and evolving very, very quickly. Incremental changes are introducing new features, new input parameters and new capabilities that make it challenging for everyone to maintain proficiency with the latest developments.
- Automation assets from GenAI systems are rarely production-ready. They can contain hallucinations and bugs. Ensure that you're validating code and establishing trust before using GenAI-created code.

Recommended Research

- Essential Skills for Infrastructure Automation Engineers
- Organize for Agility With Team Topologies
- Board Brief on Generative AI

Gather Requirements for GenAI Instrumentation, Operation and Maintenance

Organizations in every sector are investigating the potential role that GenAI functionality can play when implemented into their products and services. I&O teams should collaborate with product and development teams to gather the requirements for production GenAI implementations.

Some of these design patterns incorporate GenAI "as is" through a straightforward API call and may not require any unique infrastructure or operational processes. However, approaches that incorporate proprietary data to enhance or tailor responses, such as retrieval augmented generation (RAG), can drive requirements for specialized datastores or infrastructure, like vector databases or instances with graphics processing units (GPUs). These novel components often introduce unexpected observability gaps where the scope of the problem grows to include more than traditional performance and availability monitoring. For instance, AI-powered applications that predate GenAI usually require tailored instrumentation that measures the accuracy and relevance of model output, whether model performance is degrading and whether it continues to deliver output that is helpful, harmless and harmless. Prepare for GenAI by understanding and planning for developer and business requirements for monitoring. We expect that IT won't be the first adopter of GenAI. Rather, we expect a customer-facing or revenue-generating application to be the first cases where the learning and adoption costs can be justified. Therefore, many requirements to monitor these technologies will come from the teams that are creating those solutions.

With ever-evolving regulations on how data is stored and managed, I&O teams must also ensure these solutions adhere to regional and industry compliance standards and mandates. Collaborate with both product teams and security or legal teams to clarify and maintain these standards.

When GenAI systems are incorporated into mission-critical services, their performance and fault tolerance become paramount. Ensuring uninterrupted service for a GenAI system will change the availability problem. Now, applications, data and, potentially, models trained on that data need to be protected. Collaborate with development teams to document key components, dependencies and integration points. Evaluate and, if necessary, extend existing resilience practices to ensure I&O can protect these new assets.

Beyond standard monitoring and logging of infrastructure, focus on creating a framework for GenAI observability. This requires real-time tracking of GenAI system behavior, performance and, increasingly, performance efficiency. Observability also extends to establishing root cause analysis capabilities to identify and address issues, enhancing system reliability. In many cases, achieving observability is complicated by an inherent and important randomness in the output. This variability helps language feel more natural, but it can create scenarios where results might not be reproducible and, therefore, complicate forensic investigation into performance or behavior. Observability of GenAI not only reduces risks but also aids in understanding how the AI systems are performing and where they can be improved. Collaborate with product, development and site reliability engineer (SRE) teams to establish service-level indicators and objectives.

Early adopters are already identifying common operational requirements for delivering GenAI solutions in production. Table 3, below, summarizes their findings.

Table 1: Common Operational Requirements for GenAI

(Enlarged table in Appendix)

Requirement	Consideration
Compliance	■ Data protection and privacy laws should be considered based on the governing region of your organization. ■ Ensure that data collection, storage, processing and usage meets regulatory requirements.
Resilience	■ The production system must have the ability to handle failures and recover rapidly. ■ Underlying infrastructure should be fault tolerant. ■ Synthetic testing, load balancing and disaster recovery should all be part of the planning and design.
Scalability	■ GenAI models are computationally intensive and will require significant resources to store massive training sets for self-hosted architectures. ■ Underlying infrastructure must be capable of handling increased load on the system and be able to replicate across different regions.
Observability	■ Monitor both the performance and usage of cloud-based LLMs. ■ Modern observability of LLMs combines meaningful logging, metric collection and synthetic testing to ensure production operations.
Cost	■ Monitor the unit of licensing, usually API calls, for any LLM or GenAI implementation. This will help track and forecast costs but also help highlight architectural flaws and attacks by spotting unusual spikes or recurrent overruns.
Security	■ GenAI systems have both unique attack surfaces, like prompt injection, and attack surfaces inherent to broader AI usage, like poisoning of training data. ■ Planning for the requirements of monitoring and detecting these security scenarios might require some unfamiliar tools and skills.

Source: Gartner (September 2023)

Benefits

- Proactive investigation into GenAI and planning for its unique requirements will ensure that introducing this new technology doesn't also introduce new blind spots.
- Establishing the right kind of observability and maintenance early in the adoption will provide insight that can make the early iterations more impactful and bring success closer.

Cautions

- The patterns of GenAI usage and the longer-term goals are still emerging. Aggressive adoption will be accompanied by learning through experience. Expect to find unexpected operational challenges. If you're not prepared for iterative improvements, take a more conservative approach to adoption.
- GenAI is evolving rapidly. New features are changing both the outputs and the inputs, and backward compatibility might suffer in favor of new capabilities. This can disrupt both programmatic integration and operator skills.
- The application and data skills that support GenAI are mostly outside the purview of I&O.
- Teams solving business problems and delivering business outcomes may overlook some of the operational requirements for GenAI systems. This can cause last-minute heroics and create some technical debt to put solutions in place late in the cycle.

Recommended Research

- Launch an Effective Machine Learning Monitoring System
- AI Design Patterns for Large Language Models
- Introduce AI Observability to Supervise Generative AI

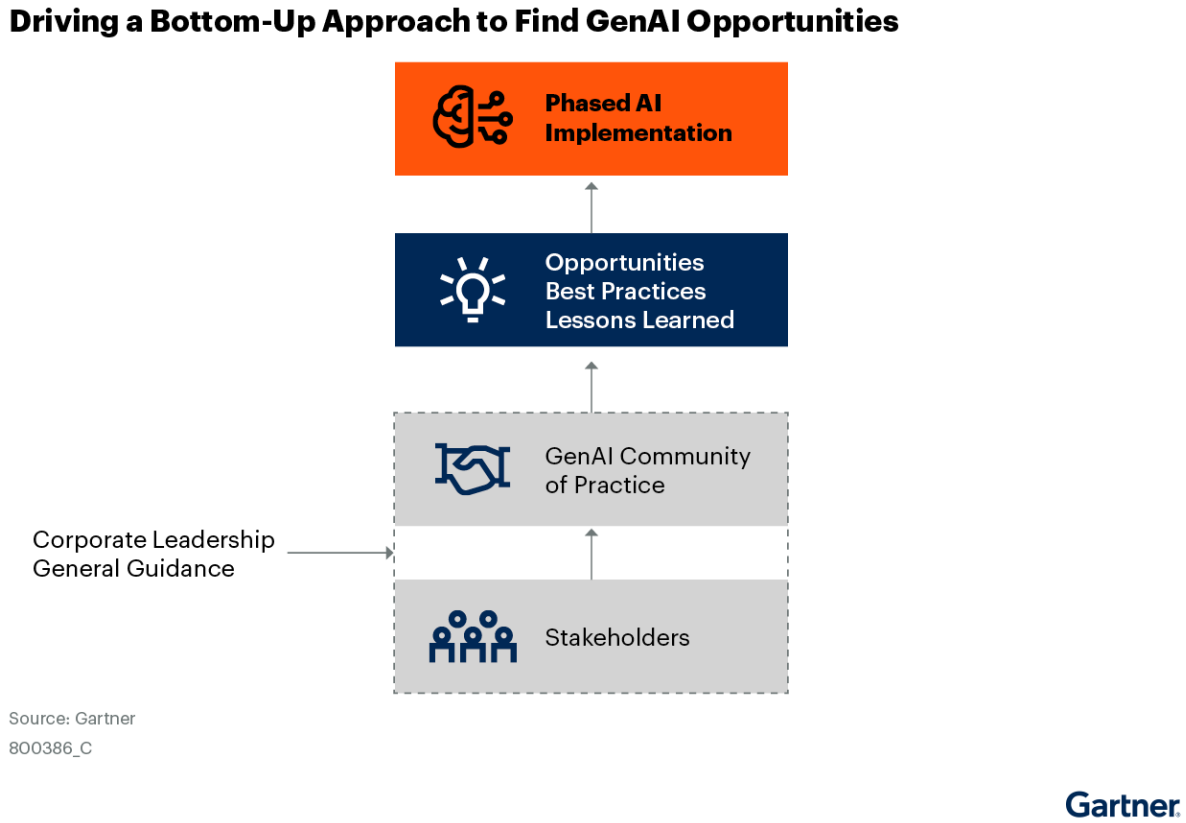
Initiate a Bottom-Up Effort to Identify Opportunities and Use Cases for GenAI

The potential of GenAI is too great and the number of use cases is too large to expect a relatively small group of leaders to identify all the top opportunities and best practices.

Although the concept of AI has been around for many years and has seen some enhancements from machine learning advancements, the launch of the OpenAI ChatGPT service has caused generative AI to take the IT industry by storm. The benefits of implementing GenAI are many. But as with any new technology, you'll need to identify opportunities, determine best practices and develop a plan to avoid common missteps when adopting the technology into a production environment.

Many organizations have chosen to accept the importance of GenAI and are looking for an effective way to adopt it. Instead of this being a top-down initiative emerging from leadership, formalize an approach that invites not just leadership but all stakeholders (I&O, DevOps, security eEngineers, etc.) to the table, as shown in Figure 3. Top-level mandates could overlook creative opportunities or even stifle efforts to implement GenAI with too narrow a perspective. Further, top-down mandates are subject to constraints, processes, RACI charts, and the like. While these will become critical elements of a production system, they shouldn't be allowed to constrain the adoption of GenAI during the ideation and exploration phase.

Figure 3: Driving a Bottom-Up Approach to Find GenAI Opportunities



For organizations to take advantage of GenAI and its adoption and implementation, a collaborative effort involving multiple stakeholders will be required, most commonly: IT leadership, software development teams, security engineers, automation engineers and I&O engineers. We recommended you take some or all of the following actions:

- Join a community of practice (CoP) to identify ways that GenAI can help your organization and how it will apply to relevant use cases. If your organization doesn't have a CoP, suggest starting one.

- Follow any formal process for submitting and reviewing ideas. Recognize creativity, encourage prototyping and participate in hackathons. Even simple proofs of concept can reveal unforeseen challenges and value, while shedding light on the realizable potential.
- Compile a list of questions that come from collaborative planning efforts (a list of RFIs). Engage with subject matter experts to find answers to these questions, such as “What are the potential candidates for GenAI implementation?”
- Learn from organizations that have implemented GenAI already, noting in particular the common mistakes that were made and how they were addressed.

Benefits

- Through grassroots efforts that bring diverse perspectives, different problems and different priorities to the table, you will uncover many opportunities and potential use cases for GenAI.
- By encouraging early grassroots brainstorming within your community of experts, you invite more creative proposals and avoid burdening every idea with the preparation required for more formal ideation.

Cautions

- Avoid a “black hole” experience where ideas are submitted and forgotten. Even when an idea is rejected outright, close the loop with the creator. This will foster goodwill and might even result in better ideas further down the road.
- At the same time, this approach can generate an overwhelming number of ideas. Manage expectations by introducing recurring phases that can time-bound the brainstorming and limit the number of immediate opportunities.

Recommended Research

- Assessing How Generative AI Can Improve Developer Experience
- Community of Practice Essentials

Increasing Demand for Self-Service Platforms Will Drive Adoption of Platform Engineering

Since the advent of public cloud IaaS, managing increasingly complex IT systems and securing valuable assets wherever they reside has only grown more difficult. Business units now often make technology purchasing decisions, independent of IT. But while purchasing power has been distributed throughout the enterprise, IT systems expertise has not. In a traditional, centralized I&O model, this frustrates both sides: Business units are upset that I&O cannot move fast enough; while I&O rightly refuses to abdicate its responsibility to keep systems reliable, available, compliant and secure.

To solve this problem, I&O technical professionals should build self-service platforms that help end users discover, operate, secure, improve and build upon complex, distributed IT systems — especially when end users are not technical experts in the underlying systems. Platform engineering is the discipline of building and operating user-centric self-service platforms, each of which is a layer created and maintained by a dedicated product team, designed to support the needs of its users. Platforms improve user experience by offering a curated set of tools and services. These are designed to present end users with best-of-breed technical capabilities and highly optimized processes, and without the need for users to create the operating platform for themselves. Platforms improve the consistency and quality of IT solutions. They reduce redundant tools and processes, consolidate parallel efforts by multiple teams, enforce security and compliance standards, and include pervasive automation.

In last year's Planning Guide, we advised I&O teams to begin building their first platforms, focusing on software developers as the initial customers and Kubernetes-based applications as the initial targets. In 2024, demand for self-service solutions will only grow. Platform engineering is once again a Gartner top strategic technology trend for 2024, and remains one of the hottest topics in DevOps (see Top Strategic Technology Trends for 2024: Platform Engineering).

To establish or extend a platform engineering practice, I&O technical professionals must:

- Manifest the evolution of DevOps culture through platform engineering.
- Adopt GitOps for stateless Kubernetes-based applications.
- Automate the testing of internal platforms to ensure performance and reliability.

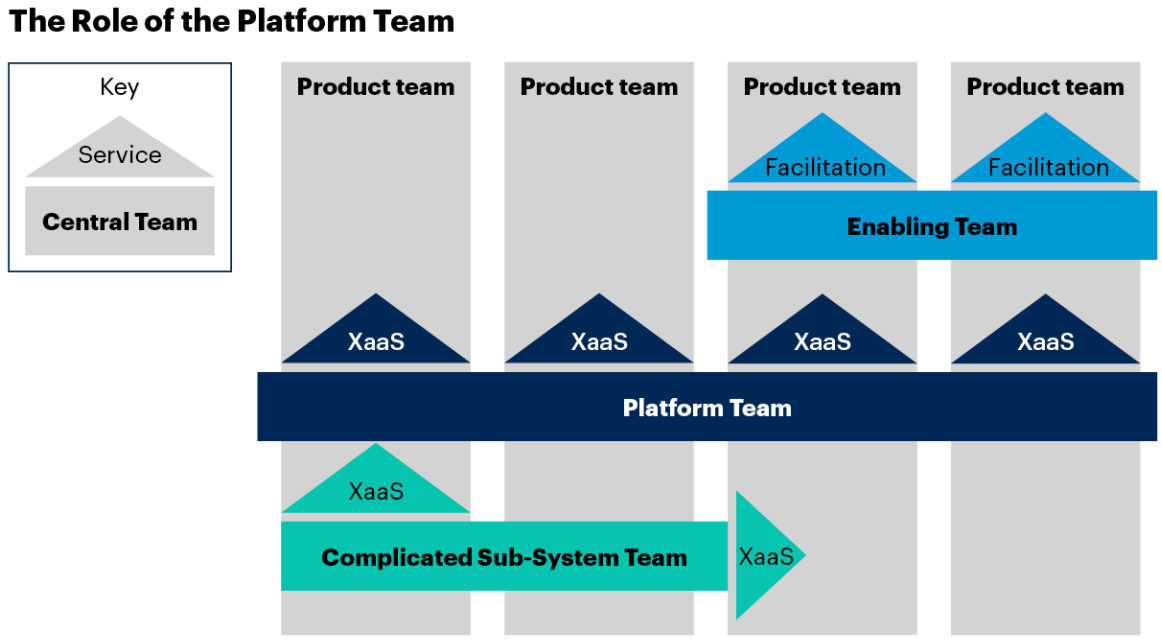
Planning Considerations

Manifest the Evolution of DevOps Culture Through Platform Engineering

The idea of closer collaboration between developers and operators is not a new one. This is the animating principle of the DevOps movement, which quickly became the standard (or at least the aspiration) of IT organizations worldwide. DevOps emerged in response to the exploding complexity of modern software, which is often built on top of an array of backing services in which developers are likely not experts. DevOps preached closer partnerships between operators, who are subject matter experts in the underlying infrastructure and services, and developers, who combine those services in useful ways. In DevOps, the role of the operator is often consultative, advising developers on how best to integrate and optimize complex, arcane services.

Platform engineering, therefore, is an extension and refinement of the original DevOps ethos. In the platform, the operators' knowledge is codified and productized. The platform is a self-contained system that has the operators' knowledge baked in, presented to developers at an appropriate level and cognitive burden. Platform engineering also defines the relationship between operations and product teams, as illustrated in Figure 5.

Figure 5: The Role of the Platform Engineering Team



Source: Gartner 2022. Adapted from Team Topologies by Matthew Skelton and Matt Pais. Used with permission.
774324_C

Gartner

In the figure, the platform spans multiple independent product teams, providing a common set of tools and architecture useful to all. The platform rolls up the work of complicated subsystems teams, such as network engineers or database administrators. These teams still exist, but their output becomes part of the platform. There may also be enabling teams that assist end users in utilizing the platform to its fullest potential.

Benefits

- User productivity is the primary benefit. The first and most important goal of platform engineering is increasing user productivity. The platform should lead directly and measurably to more work getting done by platform consumers.
- Platform engineering consolidates and centralizes the work of delivering frictionless self-service across environments and providers. Most organizations will establish dedicated platform teams to do the hard work of platform building. Whereas DevOps has a wide range of potential operating models, the general assumption is that platform teams are centralized shared service providers.

Cautions

- Lack of skills within I&O is the primary barrier. Building an internal software product is not something with which the typical I&O department has any experience. Platform teams will need deep product management and software development skills, in addition to expertise with any components or backing services on which the platform is built. Most veteran I&O personnel lack these skills and will need to build them.
- Platform engineering is the current “state of the art” for DevOps. Contrary to some vendors’ marketing claims, platform engineering does not replace DevOps; rather, it fulfills the promises of DevOps.

Recommended Research

- Top Strategic Technology Trends for 2024: Platform Engineering
- Use Platform Engineering to Scale and Accelerate DevOps Adoption
- Adopt Platform Engineering to Improve the Developer Experience

Adopt GitOps for Stateless Kubernetes-Based Applications

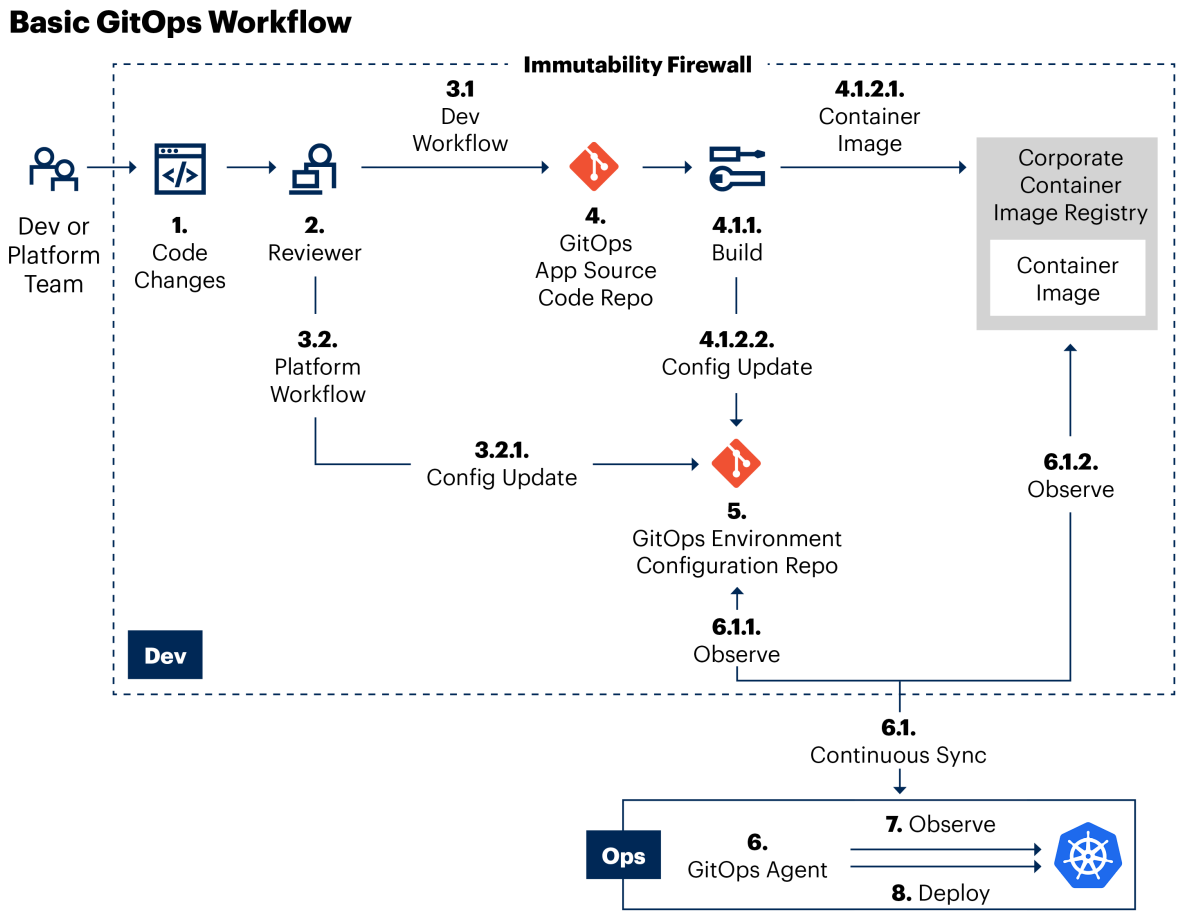
GitOps is a modern control system for cloud-native applications. It allows developers to deploy and control applications using only declarative constructs stored in a Git version control repository. GitOps creates a closed-loop, automated system of operations. The user interacts only with Git, using abstract, declarative logic. The system is self-validating and updates configuration automatically when an updated configuration is pushed into the Git repository. The K8s reconciliation loop ensures continuous configuration compliance by automatically reverting any changes not accounted for in the upstream repository. However, GitOps tools do not magically know how to convert declarative specifications into actions. There must be an operating platform that manages the underlying infrastructure and deployment pipelines, embeds security into workflows and enforces policy compliance to internal standards. GitOps and platform engineering are intrinsically linked.

Consequently, GitOps is not suitable for all use cases. Only workloads that have a suitable operating platform in place are candidates to use GitOps techniques. In practice, this means that GitOps is almost exclusively used for Kubernetes workloads, since K8s provides a substantial part of the underlying operating platform.

In 2024, I&O departments should begin the transition to GitOps, but must do so judiciously. Stateless workloads that are already running on Kubernetes make the ideal initial targets for a GitOps initiative. Your efforts should start there.

Figure 6 is a high-level reference architecture for a Kubernetes-based GitOps workflow.

Figure 6: Reference Workflow for Kubernetes-Based GitOps



The trigger point for the application deployment is when a developer commits new code to the source code application repository. Once a change to the desired state defined in the Git repository is detected by the GitOps agent, the deployment of the application through a K8s rolling update is done to satisfy the new desired state. A rolling update allows stateless application deployments to take place with zero downtime by incrementally updating Kubernetes Pods with new ones.

Benefits

- GitOps provides automated deployment and configuration management with rich potential for “as code” approaches.
- It shifts life cycle and configuration management to mature software asset management practices with reliable and extensible tools.

Cautions

- GitOps is not a panacea. It is not suitable for all workloads; and even where it is suitable, it does not come without risk and requires maturity with K8s, a robust CI/CD environment and Git integration.
- Upskill technical staff and prepare for changes. GitOps is a new way of working, with which many veteran developers and operators have no experience. Most IT organizations will need to build this expertise internally because they will be unable to hire outside talent. Kubernetes is among the most sought-after skills in IT today. Qualified candidates aren't on the job market for long, and command premium salaries.

Recommended Research

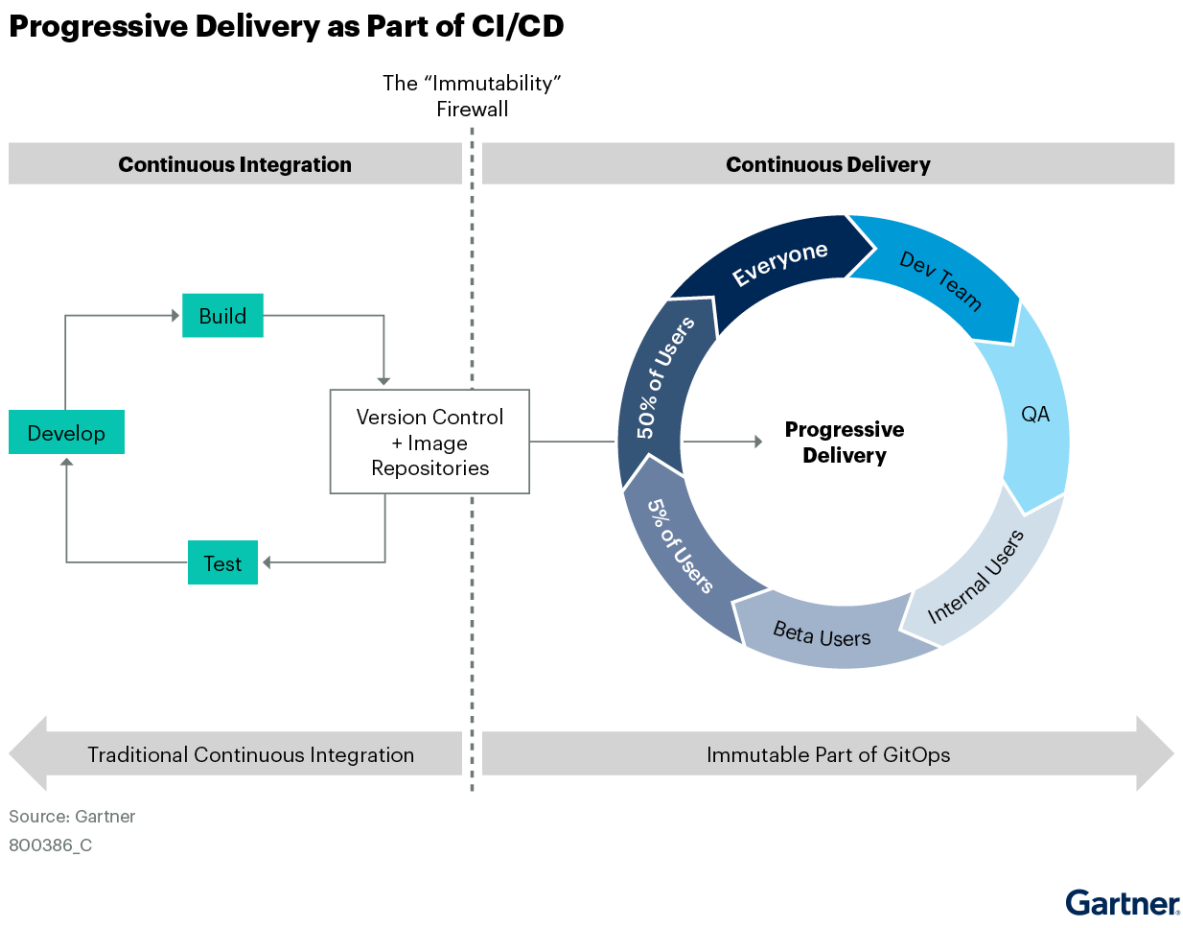
- Is Using GitOps-Style Automation With Kubernetes Right for Me?
- Innovation Insight: Top 4 Use Cases for GitOps

Automate the Testing of Internal Platforms to Ensure Performance and Reliability

Product owners looking to ensure performance and reliability of their internal platform should adopt a complete and mature software development life cycle for the platform. Every best practice for software development should be adapted and applied against the platform. This includes automation. The platform needs dedicated and automated software testing to ensure that the updates to the platform work properly, meet consumer requirements and haven't introduced any regression errors.

Some platform engineering teams rely on progressive delivery, as shown in Figure 7 below, building on the principles of CI/CD and provisioned through IaC. Progressive delivery is the key enabler of DevOps practices at scale, accounting for the need to develop automation. Practices like release progressions (development rings) and progressive delegation for exposing features into production are crucial to control performance and keep your environments reliable. Reuse your software development engineering in test (SDET) practice and discipline for IaC/policy as code (PaC) as much as possible.

Figure 7: Progressive Delivery as Part of CI/CD



Delivering this kind of automated testing can take different forms and will reflect the nature of investment around the platform. With significant investment, the testing will be commonly managed as nonfunctional requirements handled through a cloud center of excellence. Platform teams should internally have the appropriate roles and responsibilities to deal with test automation. In organizations with smaller relative investment in their platforms, the product owner should have at least some subject matter expertise in test automation, and the ability to manage the technical workspace requirements of I&O. In extreme cases, they might even own the test automation.

Benefits

- Faster feedback loops in GitOps allows product managers to minimize the potential negative impacts of failed releases.
- Progressive delivery affords an "experiment" phase between build and deploy that acts as a personal sandbox for developers. Ultimately, this provides more control over releases implemented as feature flagging, canaries and blue/green deployment.

Cautions

- Users in the progressive release phase may experience different versions of the application, depending on the features they can access. Strive to maintain consistent and coherent automation across all versions to avoid confusion or frustration.
- You need robust monitoring and logging mechanisms to track the performance of the progressive release. If any significant issues arise after the deployment, you need to be able to roll back quickly to the previous version. Having a well-defined automated rollback is essential in case of unexpected problems.

Recommended Research

- Essential Skills for Infrastructure Automation Engineers
- Automate the Application Delivery Value Stream

Differing Levels of Autonomy Will Need Different Operations Approaches

Organizations are increasingly trying to improve autonomy, which is a key element to improving user experience and thereby enabling users to do their best work. This, in turn, improves and accelerates the delivery of positive business outcomes. Greater autonomy benefits many end-user teams — not just developers. Other roles, such as data science and machine learning (DSML) teams, as well as teams responsible for managing commercial off-the-shelf (COTS) applications, are all interested in greater autonomy.

But not all teams have the same appetite or readiness for autonomy, and readiness and maturity may vary across the service life cycle. For example, an application team may be fully mature in autonomy for infrastructure provisioning, but not ready for autonomy in operations. A team's readiness will also be affected by the organization's support structures for autonomy at each life cycle stage.

These differing appetites and readiness for autonomy have begun to shift I&O functions. Technical professionals may struggle with this transition, as nearly all organizations will support a mix of legacy COTS and custom-developed applications, even if those applications are rehosted in cloud environments.

However, the roles and processes inherited from these legacy applications should not dictate the future state. Embrace the trend that even legacy applications and application owners can benefit from greater autonomy. I&O organizations must build toward a future in which application teams are offered options for autonomy, which necessitates putting support structures in place and derisking such autonomy.

In 2024, I&O technical professionals should:

- Modernize support for COTS and legacy applications.
- Establish dedicated cloud operations to maximize the value and impact of cloud adoption.
- Provide the foundations necessary for a "you build it, you run it" approach.

Planning Considerations

Modernize Support for COTS and Legacy Applications

While much of the attention around public cloud computing has revolved around net new cloud-native applications and their unique operating needs, existing applications have changing needs, too. Gartner has found that the number of existing applications that are actually refactored (or rebuilt) within organizations during public cloud migration could be lower than 30%. This means that many existing applications will move to cloud providers with no meaningful modernizations (see Quick Answer: How Much of My Application Portfolio Can I Lift and Shift to the Cloud?).

It is therefore important that organizations do not let the allure of new cloud applications act as a distraction from modernizing operations for legacy application environments.

For example, in some organizations, COTS applications may be managed by a central I&O team; in others, distinct application teams may be in charge of daily operations. Some of these applications may still end up getting moved to the cloud, especially in organizations looking to get away from running/leasing physical data center space. I&O teams must modernize their fundamental operational tools and processes, including:

- Deployment and change management
- Patching and updates
- Backup and recovery

- Audit logging
- Performance monitoring and optimization
- Inventory, audit and licensing
- Documentation and knowledge management
- Integration

Doing nothing about your operational tools while trying to modernize or migrate applications can hinder your organization in multiple ways. At a minimum, it can slow you down as the application cycle is ready to move faster than the operations/tool cycle can keep up. Worse still, the application cycle moves at full speed, leaving numerous operational deficiencies unresolved and likely creating an incident.

Organizations should therefore look to enhance (where possible) tools with many of the core attributes that make cloud computing so powerful — like end-user self-service, automation and API enablement. Whether enhancement is possible will depend largely on categories of tools (e.g., vendor-native tools, third-party tools, in-house developed tools) and the amount of effort required to improve them. Cloud initiatives are only as fast as their slowest elements, so organizations should also accelerate the improvement of tool capabilities.

Benefits

- Tool modernization will benefit many application teams throughout the organization, from traditional developers and low-code business technologists, to groups managing COTS titles and data scientists.
- Modernizing tools for a diverse array of operational environments will help organizations more quickly adopt additional cloud and edge services, thereby increasing the speed and agility of the organization.

Cautions

- I&O organizations must work with a diverse set of teams with varying skill sets to properly assess customer needs. Failure to do so — or failure to do so accurately — can lead to targeting requirements that don’t align with actual needs.

- Not all application teams may need all of the tools that have been modernized, leading to investment disparities. For example, applications that utilize a more “immutable infrastructure” deployment pattern typically would have no need for application patching tools.

Recommended Research

- Guidance Framework for Implementing Cloud Platform Operations
- Design a Data Protection Strategy for On-Premises and Cloud IaaS
- Guidance Framework for Deploying Centralized Log Monitoring
- Solution Path for Modern Infrastructure and Application Monitoring

Establish Dedicated Cloud Operations to Maximize the Value and Impact of Cloud Adoption

Operations should not be a one-size-fits-all capability, and many organizations struggle to find an approach to cloud operations that can support application teams at varying levels of cloud maturity, with different levels of cloud skills and knowledge, different paces of change, and different levels of DevOps adoption and maturity. These difficulties are exacerbated by the nature of the applications themselves, how well-suited they are to cloud environments, and the quality of their migration to the cloud.

For instance, DevOps-oriented application teams with cloud-native applications have different needs to application teams that primarily maintain custom integrations between multiple COTS applications.

While there are many possible approaches to cloud operations, Gartner recommends that organizations build a cloud platform operations function. This approach is aligned with both a service-optimized I&T operating model and the principles of platform engineering. The cloud platform operations function is effectively an internal managed service provider (MSP), with productized managed services (day-to-day operations capabilities), consulting services (project and engineering capabilities) and a software platform that supports these services with automation. Create “product lines” within the function to serve different internal customer requirements.

Implications

- The need to implement cloud operations is one of the largest drivers of I&O transformation. Although cloud operations can be embedded within existing I&O silos, this approach is almost always suboptimal. Even if the organization intends to eventually have a unified approach where cloud and on-premises operations are performed in highly similar ways, it is better to start with a dedicated cloud operations team.
- If the organization wants to eventually unify operations across environments, the best practice is to pioneer "greenfield" practices in cloud operations, and then bring those practices back to on-premises operations. Efforts to implement cloud operations identically to on-premises operations, and then gradually modernize both together, often fail, resulting in suboptimal outcomes from cloud adoption.

Benefits

- Placing cloud operations in a dedicated team focuses that team's members on cloud needs alone, on pioneering new processes and practices, and on avoiding getting mired down by legacy mindsets and existing technical debt.
- A flexible, multipronged approach to cloud operations enables I&O to meet technical end-user teams where they are, and to help drive cloud maturity at a pace that suits the needs of each team.
- Thinking about launching a cloud platform operations function, as if one were launching a startup cloud MSP, helps technical professionals to focus on how they can ramp up the team's capabilities in conjunction with gradually onboarding appropriate applications onto the team's services.

Cautions

- Some organizations embed cloud operations either in an infrastructure platform engineering team or in a platform engineering team focused on developer experience. However, this does not result in a dedicated cloud operations function, since the team has a different primary charter. While this approach can be viable in the short term — since cloud operations is often well aligned with a platform engineering approach — it usually does not scale well.

- Some organizations consider the amount of work necessary to launch a high-quality cloud platform operations function, and the skills needed, and decide that this is not an effort they can successfully execute. But many organizations can be successful with this approach if they are willing to work with a cloud MSP that can facilitate a transition through providing a platform, training and mentorship while cooperating cloud environments.

Recommended Research

- Comparing Cloud Operations Approaches
- Guidance Framework for Implementing Cloud Platform Operations

Provide the Foundations Necessary for "You Build It, You Run It"

The extreme end state of developer autonomy is "You build it, you run it" (YBI YRI). This operations approach makes application teams accept principal responsibility for the production operations of their applications, rather than relying primarily on the I&O team. Cloud-enabled developer autonomy, lack of cloud-skilled personnel in I&O, and CIO-level efforts to accelerate IT transformation have all driven increasing adoption of YBI YRI.

Successful implementation of the YBI YRI model requires:

- Application teams that are willing to embrace the model.
- Application teams that have the skills to safely and competently perform operations.
- Application teams that are already at reasonable scale, and organizational willingness to further scale these teams so they can also take on operational responsibilities.
- Secure, stable, self-service platform services that enable application teams to operate independently with minimal support or assistance from a central operations function.

Many organizations initially plan to transition to the YBI YRI model as part of a technology transition, such as during cloud migration, or for new cloud-native applications. However, just because application teams can acquire the technical means to execute the provisioning of their own infrastructure and other cloud resources, does not mean that they are ready to *operate and secure* those cloud environments.

Implementing a YBI YRI model for cloud operations, while appropriately managing the risks of that model, requires additional capabilities:

- Application teams with the necessary cloud DevOps skills to self-manage their cloud environments and resources.
- A strong foundation of adaptive governance for cloud environments and cloud users.
- Cloud-specific platform services that provide automated governance, management and security for the cloud environments — along with integration into key enterprise functions such as identity.

Implications

- Widespread adoption of YBI YRI within an organization radically changes the role of the I&O organization in ways that are generally beyond the control of technical professionals. I&O technical professionals in such organizations should carefully consider the future roles that will be available to them, changes to their existing roles and the required skills, and plan accordingly.
- Organizations must possess excellent platform engineering capabilities, with a strong emphasis on developer experience, to deliver the necessary automation scaffolding to support YBI YRI. The high degree of automatability present in the public cloud services delivered by market-leading providers eases these efforts. However, organizations cannot simply buy a cloud platform “off the shelf.” Instead, they often spend considerable effort integrating multiple commercial products and developing additional automation themselves.

Benefits

- YBI YRI can help accelerate the release velocity of application teams that are willing to accept greater operational responsibilities. This approach can be particularly attractive in the cloud — especially for architectures that primarily use PaaS — because of the greatly reduced need to manage infrastructure.
- YBI YRI is often most beneficial when coupled with a site reliability engineering (SRE) philosophy, with SREs either embedded in application teams or in the software engineering function as a whole. This approach helps place the operational focus on delivery against service-level objectives, and discourages the accumulation of technical debt.

- YBI YRI can help alleviate some of the pressure on I&O teams that are experiencing a shortage of staff with appropriate cloud and DevOps skills. This can be especially useful when the software engineering function has an abundant budget but I&O does not.

Cautions

- YBI YRI is a very common operations approach for start-ups and digital-native teams piloting applications in the cloud. It evolves naturally because the I&O team doesn't yet have cloud skills to formally support the pilot and, therefore, finds itself willing to abdicate its responsibilities to the pilot teams. The pilot teams, on the other hand, have few guardrails and are willing to do whatever it takes to deliver the desired business outcomes. As quickly as possible, establish the skills necessary to address the governance gap and make this transition a prerequisite for moving from pilot to formal adoption.
- Your organization may be able to force application teams with cloud-based applications to use a YBI YRI model regardless of their willingness to do so. This might not be immediately disastrous. However, it is likely to result in mounting risk and accidental technical debt that will eventually have consequences.

Recommended Research

- Introducing the "You Build It, You Run It" Modern Operations Pattern
- How to Empower Technical Teams Through Self-Service Public Cloud IaaS and PaaS
- Solution Path for New Roles and Skills to Develop Your Future I&O Career

Shifting Accountability Will Require Versatility and Collaboration

Increasingly, IT is focusing on higher-level business objectives and driving the need for cross-functional, symbiotic relationships between stakeholders accountable for achieving these outcomes. Accountability for many of these business outcomes will transcend organizational units. Success requires cross-team collaboration between teams that don't usually work together, integrating their expertise and processes.

These teams require tools that enable collaboration without disrupting efficiency, effectiveness and flow. While integration between tools will facilitate these cross-departmental workflows, it is the role of the versatilist — those with deep expertise in a variety of skills — to bridge the gap between teams and bring them together for a common purpose:

- **For I&O technical professionals and cloud architects**, this means collaborating to improve the accountability within a self-service environment. Consumer autonomy needs to bring considerations of security, costs and resilience into the design process. Cost overruns, for instance, need to be addressed before the money is spent. Foster a culture in which data is shared with cloud consumers to enable their direct accountability, and make them aware of how their consumption and design decisions impact operational overhead.
- **For SecOps and monitoring teams**, this means identifying and sharing telemetry that improves end-to-end visibility for both security and operational processes. The increased visibility will improve anomaly and threat detection to better safeguard the organization, and improve the availability of the digital service on which the organization depends.
- **For IT service management (ITSM) teams and developer teams practicing DevOps**, this means gaining a better understanding of each team's motivators and modern capabilities, then working together to strike a better balance between costs, agility and reliability.

I&O technical professionals should take the following actions in 2024 to begin shifting to a shared accountability model for the co-delivery of digital products and experiences to consumers.

- Shift accountability for cloud expenses to self-service consumers.
- Unify observability and security through expanding monitoring tool capabilities.
- Improve the flow of value by integrating ITSM and development workflows.

Planning Considerations

Shift Accountability for Cloud Expenses to Self-Service Consumers

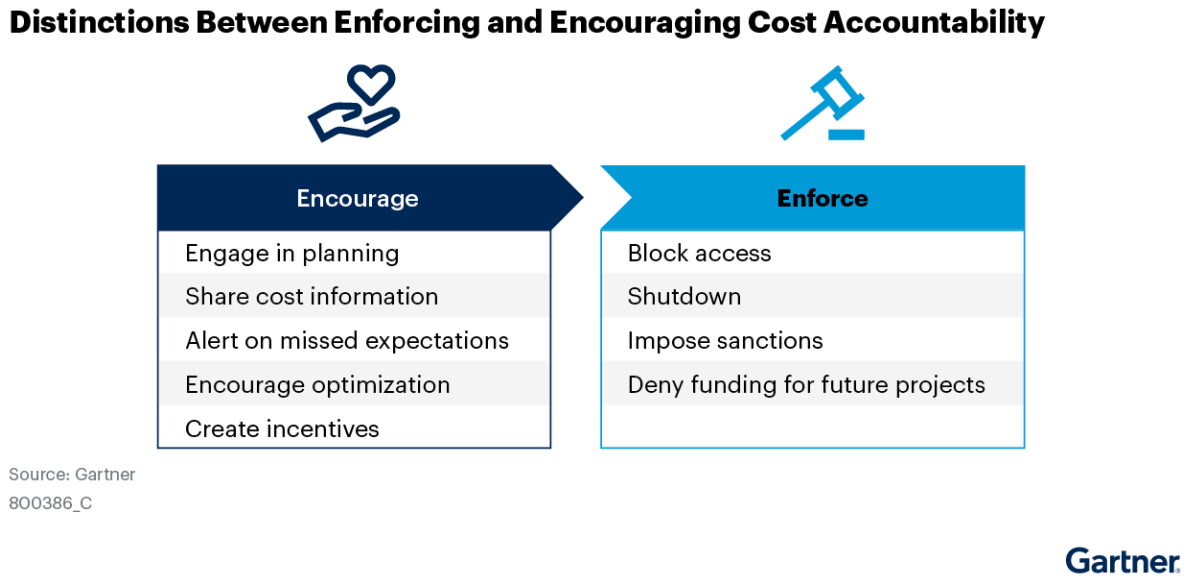
In light of the shift to platforms and self-service, organizations must redefine the accountability model of IT service expenses. Self-service means that IT consumers are empowered to make decisions about service consumption without consulting I&O or a cloud platform team. They select the architectural patterns, service configurations, resource number and size required to accomplish a business outcome. However, because cloud costs are accrued based on consumption metrics, these decisions create billable liabilities for your organization. The only way to manage these liabilities without compromising the self-service model is to hold the consumers accountable for their cloud consumption.

Shift the accountability of cloud expenses to cloud consumers by making them responsible for respecting application budgets. Due to the cultural implications of budget accountability, the shift to cloud consumers cannot happen overnight. You must encourage this shift by implementing a cloud financial management (CFM) capability that provides cloud consumers with the right information to take ownership of their cloud expenses. To shift budget accountability, CFM demands you to:

- Engage cloud consumers in cloud planning activities.
- Share cloud cost information with cloud consumers.
- Alert cloud consumers when costs are missing expectations.
- Encourage cloud consumers to optimize their deployments.
- Implement incentives to increase the involvement of cloud consumers in managing costs.

Ultimately, a sense of accountability will encourage more cost-conscious consumption of cloud services and more cost-efficient architectures (see Figure 8).

Figure 8: Distinctions Between Enforcing and Encouraging Cost Accountability



Traditional on-premises IT was based on a centralized budgeting approach. Such a model was effective because the IT organization was part of any IT project and could, therefore, exercise control over resource consumption. Cloud computing has reversed this paradigm, with IT now often involved too late when key spending decisions have already been made by other teams. Even if no longer in full control, shifting cloud expense accountability enables I&O technical professionals to influence cloud spending decisions and ensure they're made in consideration of benefits, risks and constraints.

Once the budgets are in place, enforce them by defining clear consequences when budgets are not respected. Consequences may include access blocking, automated shutdowns, disciplinary sanctions or simply an increased difficulty in getting approvals for future projects. Wherever applications are delivered through a mature CI/CD pipeline, shift left by automating the assessment and enforcement of budgets through policy as code (PaC).

Benefits

- Cost-conscious cloud consumption reduces overspending and waste, which can threaten the success of cloud adoption initiatives.
- Accountable cloud-consuming teams can better align cloud spending with business value, increasing the gross margin of digital business initiatives.
- A pervasive cost-conscious culture enables more scalable use of cloud computing while managing financial risks.

Cautions

- A traditional culture of centralized budget and predictable IT spending may be difficult to displace in favor of a more agile and decentralized spending model.
- Cloud-consuming teams may not be able to prioritize cost-efficiency when they are also asked to deliver high-performing business capabilities fast and securely.
- Large-revenue business lines may not justify an investment in this shifted accountability model or in the overall CFM capability.

Recommended Research

- Beyond FinOps: The Gartner Framework for Public Cloud Financial Management
- Beyond FinOps: Budgeting and Forecasting Public Cloud Costs
- Beyond FinOps: Gaining Visibility Into Your Public Cloud Costs
- Beyond FinOps: Optimizing Your Public Cloud Costs
- Beyond FinOps: Empowering Cloud Consumers to Manage Their Public Cloud Costs

Unify Observability and Security Through Expanding Monitoring Tool Capabilities

For every organization, security is a primary concern. Traditionally, security monitoring and application performance monitoring (APM) have been separate initiatives, with separate tools and separate teams. The roles and tools were clear: security monitoring was the responsibility of security organizations (e.g., infosec) and APM was the responsibility of I&O and development teams.

Although security monitoring and APM have had different end goals, both depend on a common set of telemetry. APM and observability vendors are capitalizing on this shared data and are integrating security capabilities into their solutions to support cross-organizational teams and functions (see Figure 9).

Figure 9: Converging Roles Within Observability Solutions

I&O technical professionals responsible for APM and a shared model for application security must:

- **Partner with security teams in documenting existing application security monitoring gaps:**
  - Identify opportunities to leverage telemetry data — unique to software observability solutions — to detect application security vulnerabilities missed by traditional security monitoring tools.
  - Combine security and application expertise on a unified set of tools and processes to improve your overall security posture.
- **Explore and leverage the security capabilities of existing observability tools:**
  - Collaborate with your security principals and vendors to evaluate whether the security features of your observability tool are fit for purpose and, if they are, whether they are beneficial or redundant to existing solutions. Establish metrics to assess functional gaps and align on processes for human collaboration to address short-term gaps.
  - Leverage the additional capabilities of your existing software observability solution for a cost-effective and timely implementation.

- **Generate security alerts and build role-based views within your observability solution.**
  - Partner with security to establish the right security alerts and how to route those alerts.
  - Take advantage of custom dashboard features to build tailored views that provide stakeholders with actionable data specific to each role.

Benefits

- A stronger culture of collaboration using shared tools and data will enable deeper expertise to be brought to bear on issues.
- Implementing an APM or observability solution will mean fewer agents, resulting in lower administrative effort and reduced resource consumption.
- Runtime application self-protection (RASP) can enhance existing security monitoring by detecting application vulnerabilities and preventing threats at runtime.

Cautions

- Obtain agreement and consensus at all levels when implementing an APM or observability solution for security use cases. Ensure there is clear ownership of costs and agreements on data access. Organizational barriers are not easily surmounted, and APM or observability solutions may garner a negative reputation if not effectively deployed.
- Clearly define roles and responsibilities for the routing of and response to security vulnerabilities and threats. APM or observability solutions provide new security capabilities and telemetry data. Organizations need to be prepared to successfully respond to the data collected.

Recommended Research

- Assessing OpenTelemetry's Importance to Application Performance Monitoring
- Critical Capabilities for Application Performance Monitoring and Observability
- Magic Quadrant for Application Performance Monitoring and Observability

Improve the Flow of Value by Integrating IT Service Management and Development Workflows

The relentless demands to increase the development speed, shorten the time between deployments, and ensure service quality and reliability are universal challenges for I&O teams.

Regardless of your development or operations models, your pace of digital innovation is a shared responsibility between development teams and operations teams. But many of these teams are guided by stand-alone performance indicators that measure the output of their individual efforts, which often leads to counter-productive working relationships between teams. These counter-productive relationships form when product teams, software engineering leaders and application leaders choose to measure their teams' success solely on deployment frequency, while I&O leaders myopically focus on uptime metrics.

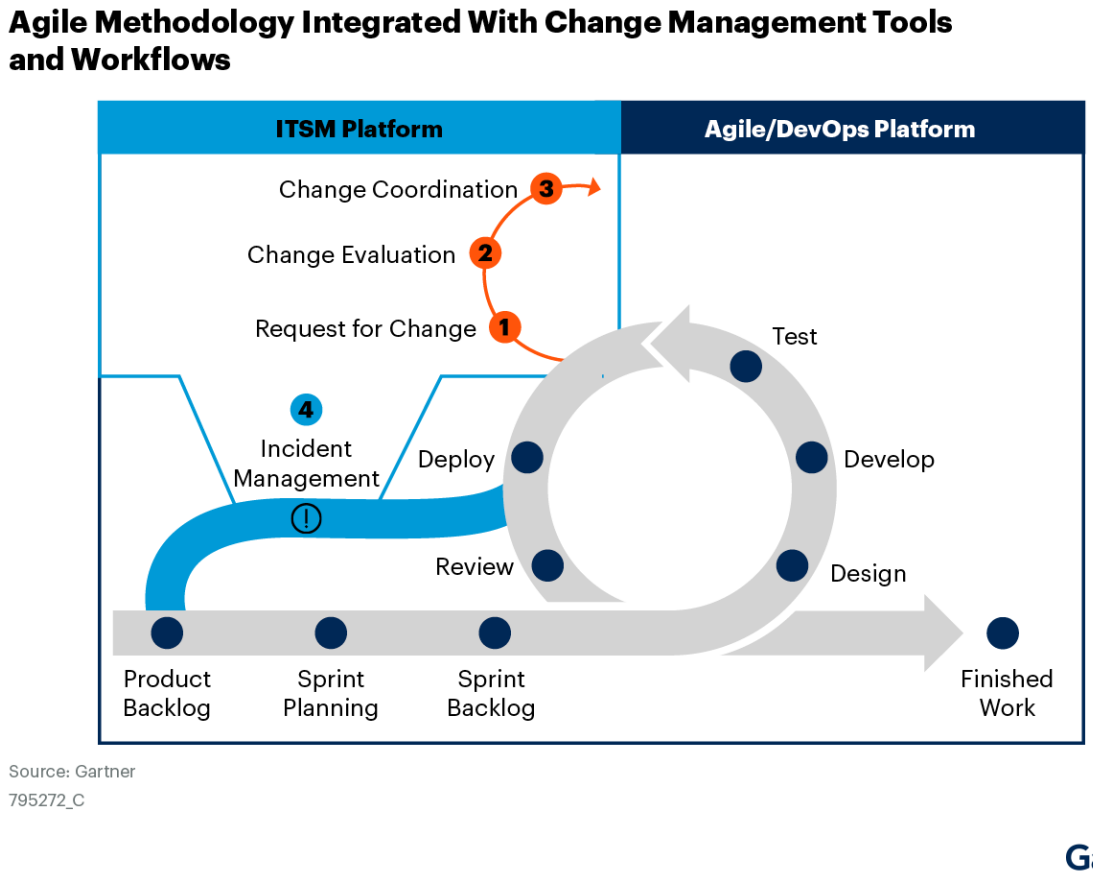
Past attempts to reconcile these differences were often thwarted by the belief that increased speed came at the cost of reduced accuracy, and vice versa. While the speed-accuracy trade-off is a limiting factor for manual tasks, it is not a constraint for modern platforms capable of integrating and automating many development and ITSM workflows.

In 2024, I&O technical professionals should shift to a shared accountability model to co-deliver digital products and experience by:

- Aligning I&O and development teams to a common set of business outcomes and sharing accountability for achieving those outcomes across their teams.
- Integrate I&O and development teams, tools and workflows to improve efficiency, effectiveness and flow.
- Leverage modern ITSM and DevOps platform capabilities to improve agility and reliability without trading one for the other.

The key integration and modernization points to target in 2024 are highlighted in Figure 10.

Figure 10: Agile Methodology Integrated With Change Management Tools and Workflows



Benefits

- Autocreating change requests will improve the developer and operator experience by enabling them to remain in their respective specialized platforms. This integration will improve the flow of work across teams to improve their efficiency.
- Automating change evaluation will help identify which changes to autoapprove and which changes need additional scrutiny to reduce the demand on I&O resources while increasing change velocity.
- Automating change coordination will help detect planned and unplanned events that may prevent the release from being successfully deployed, and reduce service interruptions.
- Integrating change and incident management will help identify issues caused by change to improve incident response time and collaboration to improve service reliability.

Cautions

- Resist the urge to begin this journey without first agreeing on the destination (outcomes). Define where you are going, how you will measure your team's progress toward that destination and who will be accountable for getting you there before changing tools and workflows.
- If the methodology your team is following is so rigid or closed off that it is unable to integrate with and complement other methodologies, it is no longer a viable way of working. Your teams will also become increasingly isolated from the rest of the organization. If this sounds familiar, it is likely not the methodology itself, but rather the way the methodology has been interpreted and implemented in your organization. Refocus your team on finding common ground between methodologies to achieve shared outcomes.
- You may have already acquired some or all of the capabilities highlighted in Figure 10. If this is the case, it is time to start using them. However, if your ITSM platform does not have these capabilities, or if the cost of these capabilities within your platform is too high, it is time to start exploring alternative platforms that meet your organization's needs.

Recommended Research

- ITSM Best Practices: How to Implement an Effective IT Change Management Practice
- How to Harden IT Process Automation Security Practices
- ITSM Best Practices: Automating Incident Management

Resilience and Continuity Will Become Strategic Imperatives Across Operations

As an organization matures its cloud usage, there is often a realization that limiting the role of cloud to a hosting location or an infrastructure as a service (IaaS) provider misses most of the value. The impact of cloud on resilience is disruptive, but successfully capturing the value of that disruption requires adapting approaches to resilience and continuity. There's no magical or effortless way to adapt; lifting and shifting applications to the cloud will not transform the resilience of an application.

In modern cloud environments, you must avoid failures by establishing built-in resilience with critical systems. Less-critical systems, lacking built-in resilience, still require disaster recovery (DR) models. It is crucial, therefore, to devise a robust cyber-resilience strategy for all workloads to counter cyberthreats such as ransomware beyond the scope of existing resilient and DR frameworks. The unifying thread is automation, knitting system components into a seamless fabric of end-to-end efficiency, featuring cost-effective, high-quality and repeatable processes.

In 2024, to complete strategic operations, I&O technical professionals need to:

- Expand protection of valuable data to address the risks of cyberthreats.
- Embrace multifaceted resilience to provide adaptive protection across hybrid environments.
- Design automation for resilience.

Planning Considerations

Expand Protection of Valuable Data to Address the Risks of Cyberthreats

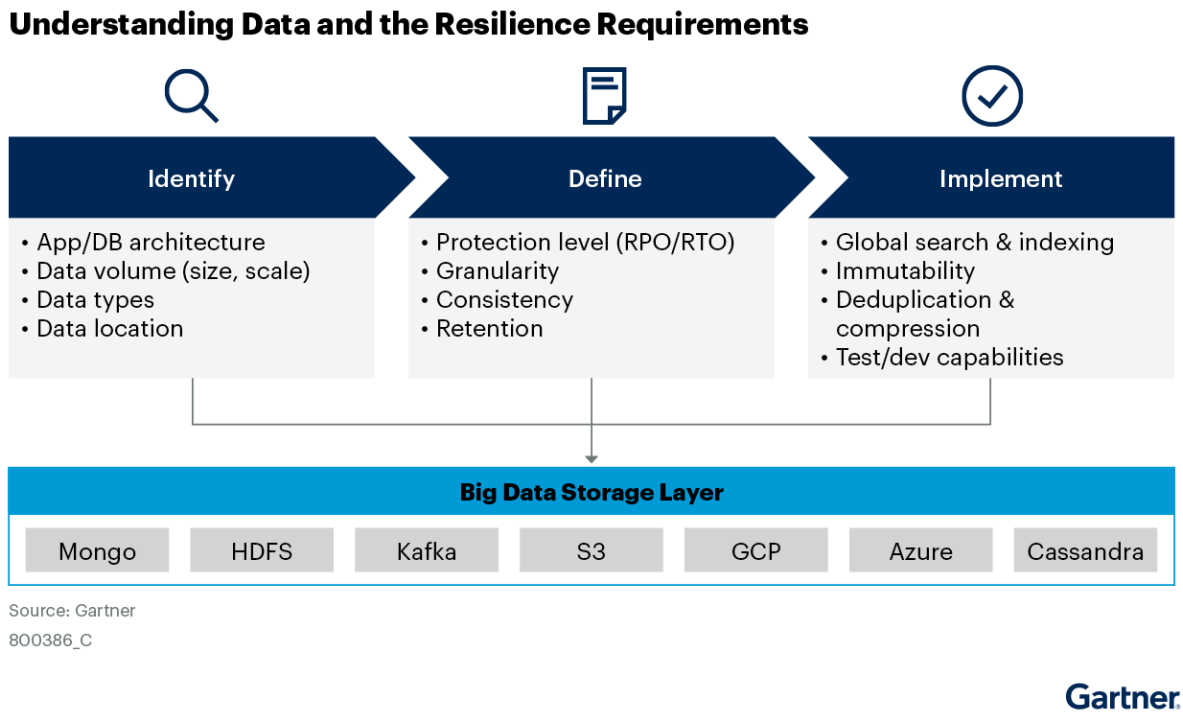
The term “cyber resilience” has been around for years and many organizations have achieved very mature practices to protect their environments. However, with the rise of ransomware attacks, that protection now needs to extend beyond servers and infrastructure. Improve your cyber resilience by adopting modern data protection, especially for regulated industries such as healthcare, financial services and manufacturing.

Data growth has been unprecedented and, with the advent of AI and its dependency on data, the growth trend is only going to accelerate.

Modern applications are generating and consuming data at a blistering pace. All forward-thinking organizations will begin to leverage large data lakes as part of AI/ML architectures. As this data continues to grow across various applications, clouds and platforms, so too does the attack surface for cyberthreats such as ransomware. Resilience in the form of (storage) snapshots or replication will not mitigate (sufficiently) against these threats, nor that of accidental deletion or internal bad actors.

It is often said that data is the lifeblood and foundation of an organization. To that end, organizations must operationalize a data protection strategy that expands into modern application architecture and ensures recoverability, availability and integrity of mission-critical data. The framework for such a strategy is outlined in Figure 11, which shows data sources that must be considered as part of an organization's cyber resilience challenge.

Figure 11: Understanding Data and the Resilience Requirements



The exponential growth of data persistence in modern applications will force application owners and data protection teams to collaborate in designing data protection strategies that can adapt to the scale and speed of change required. The data protection requirements will scale to the petabyte (PB) range, with recovery point objective (RPO) requirements close to real time becoming a requirement. These large datasets will require near-instant recovery, and that will mean exploring radically different toolsets and approaches to data protection than what have been used in the past.

Benefits

- **Comprehensive protection:** Having secure backup data stored outside the production platform, combined with immutability and encryption, ensures organizations have defense in depth against ransomware and other data exfiltration attacks, which have become increasingly common and frightening.

- **Enhanced recovery options:** A backup and recovery strategy for data within a data lake such as Amazon S3 or Microsoft Azure Blob Storage will provide much more comprehensive options for recovery than resilient options such as snapshots or replicas. The ability to search through large volumes of data and restore specific data granularly with speed will be a benefit that organizations will not be able to ignore and, for compliance reasons, most likely require.
- **Life cycle and retention management:** Many data lakes may only require limited retention for operational recoveries, but compliance-sensitive information may need to be retained for several years. A comprehensive strategy for this data will provide flexible policy-driven options for managing data retention requirements.

Cautions

- **Cloud cost optimization:** Data used for modern applications leveraging AI will inevitably be hosted on public cloud storage. Amazon S3 and Azure Blob Storage are key components of cloud data lake architecture. Organizations need to be cautious of spiraling cloud costs that will stem from extended retention and egress charges.
- **Multicloud management:** Modern applications may consist of data from multiple platforms (polyglot persistence), all with different options for data protection. This can lead to complex data protection architectures that require additional management and increase the total cost of ownership.

Recommended Research

- Ransomware Recovery Requires a Layered Recovery Response

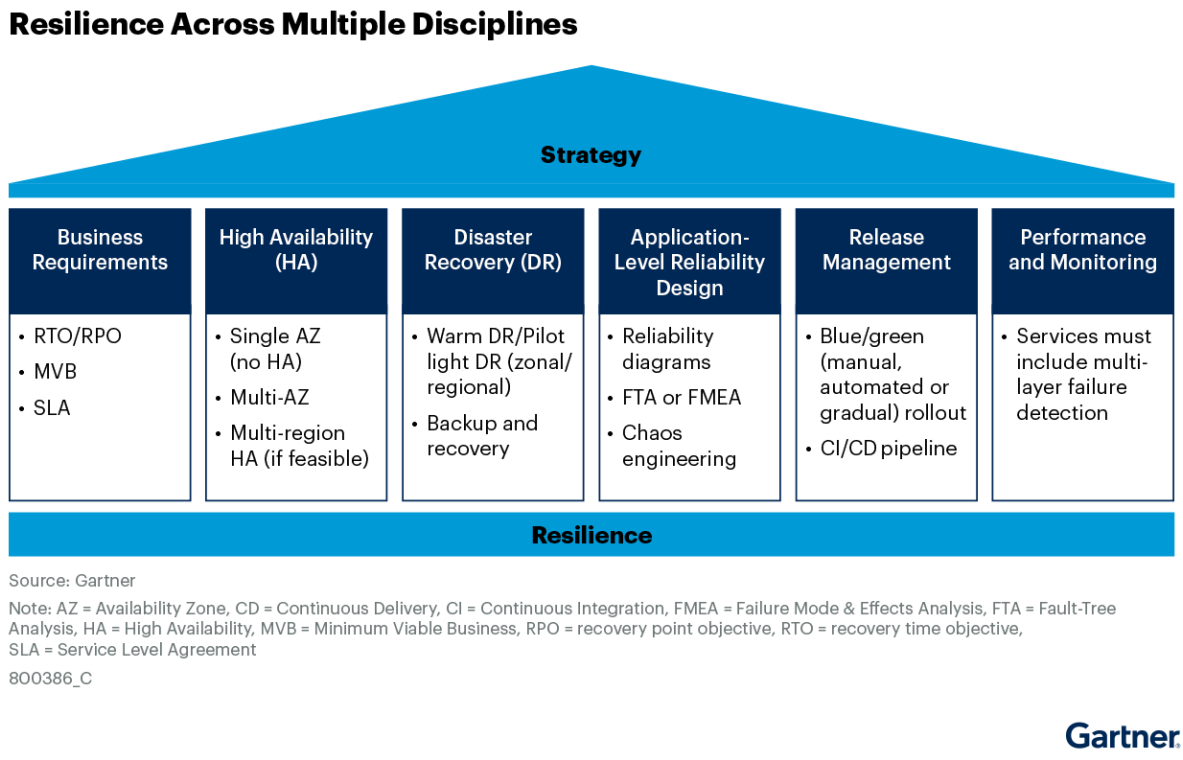
Embrace Multifaceted Resilience to Provide Adaptive Protection Across Hybrid Environments

Multifaceted resilience ensures that design, build and implementation phases all take into account a spectrum of resilience strategies. It is not a single strategy that will ultimately facilitate operational continuity. The goal is to make resilience the core foundational element to maintain organizational continuity. It starts with the belief that it is always preferable to sidestep failure whenever possible. This is achieved by improving the factors that make up resilience, thus improving overall operational capabilities and efficiency.

I&O technical professionals must ask themselves the question: "How can organizations keep critical applications, services and data available, and limit downtime and impact during failure events?" The answer is implementing continuously available, mission-critical applications and integrating technical architecture for DR into the application architecture processes.

Cloud platforms offer a variety of services and form factors beyond instances/virtual machines (VMs) for applications that are not optimized to run in traditional IaaS based on VMs. To embrace multifaceted resilience, the practice of moving largely unchanged workloads to the cloud must be exchanged for the standards of a foundational resilience approach, as represented in Figure 12. Obviously, rearchitecting every application migrated to the cloud isn't going to work. But don't overlook where the nature of cloud can afford additional layers and approaches to resilience that don't require reworking applications. For example, multiregion or multiavailability zone (AZ) deployments can improve the resiliency of a clustered application.

Figure 12: Delivering Resilience Across Multiple Disciplines



Benefits

- Limiting the impact of failures and recovering quickly from failures prevents costly downtime.

- Reducing the amount of reworking or retrofitting applications to attempt to make them resilient.
- Grouping of workloads and repeatable processes leads to opportunities to reduce or remove hidden costs and complexity.

Cautions

- Early decisions will have the biggest impact on the eventual outcome of your solutions and may also be the hardest ones to reverse later. Foundational resilience forces requirements that can positively impact availability, capacity and flexibility.
- Feasibility of resilience is dependent on the proper identification of critical workloads. Many organizations can prioritize workloads at too high of a criticality, making it difficult to achieve any of the targets for resilience.
- Infrastructure teams often drive DR and when an infrastructure team is pushing resilience applications without the direction of application architecture, the resilience is an afterthought.
- Bugs will happen. You can't eliminate them, but you can survive them. Resilient application architecture allows you to survive myriad cloud failures, including bugs.

Recommended Research

- How to Build a Secure Environment to Recover From Ransomware and Other Cyberattacks
- Use Adversary-Generated Threat Intelligence to Improve Threat Detection and Response

Design Automation for Resilience

Automation requires decomposing outcomes into a series of mechanical and repeatable steps. Usually, automation increases reliability, but planning for automation failures and building automation resilience are often overlooked. When a rare automation failure happens, troubleshooting and recovering are complicated by the fact that the process or system may be in an unexpected or unfamiliar state. This can slow and extend the efforts to shift to human effort or restore automation.

But it isn't possible or cost-effective to build an automation process that is incapable of failing. Even if the automation code is flawless, failures can still result from external forces or latent security vulnerabilities in the automation itself. Therefore, the only way to eliminate this risk is to design automation systems that are resilient to the types and impact of automation failures.

Automation-enabling resilience takes many forms. Make the automation process resilient with robust exception handling to address common issues. Preserve original data for the duration of the automated operation to enable instant rollback to a known-good state in the event of an exception. Document and rehearse manual execution modes for critical processes. Adopt development and security best practices to avoid and eliminate vulnerabilities found in shared libraries. All of these resilience techniques — individually and collectively — allow for more ambitious use of automation.

Support recovery by building automated processes that attempt to fail gracefully, capturing state and progress before terminating when possible. Consider the pattern of integrating an automated handoff within critical automations for unhandled exceptions. In this case, the automation might send a notification or create and assign a ticket. Use that automated communication to initiate the handoff to manual processing, inform the team of the state of the process or system, and to explain why the automation failed.

Benefits

- Automation resilience reduces the risk of automation failure and minimizes disruption when a failure occurs.
- It allows human operators to gracefully assume control of tasks and processes with an understanding of the system state and why the system is in this state.
- It increases the applicability of automation to more critical and more complex systems where trust and reliability are uncompromising.

Cautions

- Not all automated processes will merit the effort of designing them with layers of resilience. Adopt a system mindset and map how automation enables or supports other systems and processes. Define failure modes for these dependencies and iteratively improve the resilience of the system, not just individual automations.

- These approaches can raise the automation skill bar even higher. I&O automation skills are always in high demand and building systems of resilient automation will make the challenge of finding personnel with the required skills even greater. Where possible, create standard exception handling and notification functions that can be incorporated into other automations.

Recommended Research

- How to Harden IT Process Automation Security Practices
- To Automate Your Automation, Apply Agile and DevOps Practices to Infrastructure and Operations
- Automate the Application Delivery Value Stream

Evidence

[1] Challenges and Applications of Large Language Models, arXiv, Cornell University.

Note 1: Large Language Models

Large language models (LLMs) generate intrinsic and extrinsic hallucinations. Extrinsic hallucinations mean the generated text logically contradicts the source content. Intrinsic hallucinations mean the output is under-determined; we cannot verify the output correctness from the provided source.

Document Revision History

2023 Planning Guide for IT Operations and Cloud Management - 10 October 2022

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

2024 Planning Guide for Security

2024 Planning Guide for Data Management

2024 Planning Guide for Software Development

2024 Planning Guide for Cloud, Data Center and Edge Infrastructure

2024 Planning Guide for Identity and Access Management

2024 Planning Guide for Application Architecture, Integration and Platforms

Table 1: Common Operational Requirements for GenAI

Requirement	Consideration
Compliance	■ Data protection and privacy laws should be considered based on the governing region of your organization. ■ Ensure that data collection, storage, processing and usage meets regulatory requirements.
Resilience	■ The production system must have the ability to handle failures and recover rapidly. ■ Underlying infrastructure should be fault tolerant. ■ Synthetic testing, load balancing and disaster recovery should all be part of the planning and design.
Scalability	■ GenAI models are computationally intensive and will require significant resources to store massive training sets for self-hosted architectures. ■ Underlying infrastructure must be capable of handling increased load on the system and be able to replicate across different regions.

Requirement	*Consideration*
Observability	■ Monitor both the performance and usage of cloud-based LLMs. ■ Modern observability of LLMs combines meaningful logging, metric collection and synthetic testing to ensure production operations.
Cost	■ Monitor the unit of licensing, usually API calls, for any LLM or GenAI implementation. This will help track and forecast costs but also help highlight architectural flaws and attacks by spotting unusual spikes or recurrent overruns.
Security	■ GenAI systems have both unique attack surfaces, like prompt injection, and attack surfaces inherent to broader AI usage, like poisoning of training data. ■ Planning for the requirements of monitoring and detecting these security scenarios might require some unfamiliar tools and skills.

Source: Gartner (September 2023)