Hype Cycle for Emerging Technologies, 2023

Published 2 August 2023 - ID G00793566 - 103 min read

By Analyst(s): Arun Chandrasekaran, Melissa Davis

Initiatives: Digital Future; CIO Technology and Innovation Leadership

The disruptive technologies in this Hype Cycle will affect business and society through 2033. Technology innovation leaders, including CTOs, can use them to harness emergent AI, enhance developer experience, exploit the pervasive cloud, and deliver human-centric security and privacy programs.

More on This Topic

This is part of an in-depth collection of research. See the collection:

- 2023 Hype Cycles: Deglobalization, AI at the Cusp and Operational Sustainability

Analysis

What You Need to Know

The emerging technologies on our Hype Cycle fall into four themes:

- **Emergent AI:** The popularity of many new AI techniques will have a profound impact on business and society.
- **Developer experience (DevX):** Developers must have good all-round interactions when creating software and services if digital initiatives are to be successful.
- **Pervasive cloud:** Cloud computing will drive business innovation as it becomes more distributed and focused on vertical industries.
- **Human-centric security and privacy:** Organizations will become more resilient by using security and privacy techniques that create a culture of mutual trust and awareness of shared risks between teams.

As a technology innovation leader — perhaps a CTO — you must follow emerging technologies and applied frameworks to determine their impact on your industry and the opportunities for your organization.

Use this Hype Cycle to:

- Evaluate the business impact of emerging technologies and learn Gartner's recommendations for how to use them to drive competitive differentiation and efficiency.
- Examine technologies with transformational potential for your business and technology capabilities, exploring how you could use them for various use cases.
- Strategize how to exploit these technologies in line with your organization's ability to handle unproven technologies.

The Hype Cycle

This Hype Cycle is unique because it distills insights from more than 2,000 technologies and applied frameworks that Gartner profiles each year into a succinct set of "must-know" emerging technologies. We comprehensively assessed and analyzed both Gartner internal and external data sources to select technologies for their potential transformational benefits and their broad impact.

The technologies in this Hype Cycle are at an early or embryonic stage. Great uncertainty exists about how they will evolve. Such embryonic technologies present greater risks for deployment, but potentially greater benefits for early adopters.

This Hype Cycle typically introduces technologies that haven't featured in previous iterations. Limited space means we've removed many technologies highlighted in the 2022 version. Those remain important, and most are featured in other Hype Cycles (see the Off the Hype Cycle section).

Themes in Emerging Technologies

In 2023, our emerging technology coverage focuses on emergent AI, developer experience, pervasive cloud, and human-centric security and privacy.

Emergent AI: The massive pretraining and scale of AI foundation models such as GPT-4, viral adoption of conversational agents such as ChatGPT and the proliferation of generative AI applications are heralding a new wave of workforce productivity and machine creativity. Although generative AI possesses much potential for technology innovation, several other emerging AI techniques also offer immense potential for enhancing digital customer experiences, making better business decisions and building sustainable competitive differentiation.

Evaluate:

- AI simulation
- Causal AI
- Federated machine learning
- Generative AI
- Graph data science
- Neuro-symbolic AI
- Reinforcement learning

Developer experience: Enhancing DevX is critical for most enterprises' digital initiative success. It's also vital for attracting and retaining top engineering talent, which companies currently find difficult, and for ensuring that team morale remains high and that work is motivating and rewarding. DevX refers to all aspects of interactions between developers and the tools, platforms, processes and people they work with to develop and deliver software products and services.

Examine:

- AI-augmented software engineering
- API-centric SaaS
- GitOps
- Internal developer portal
- Open-source program office
- Value stream management platforms

Pervasive cloud: Over the next 10 years, cloud computing will evolve from a technology innovation platform to become pervasive and an important driver of business innovation. To enable this pervasive adoption, cloud computing is becoming more distributed and focused on vertical industries. Cloud providers are offering cloud services outside public cloud regions (e.g., on-premises) and reimagining the cloud at the edge, making it more vertically integrated. Data sovereignty needs are also driving distributed cloud computing. Industry cloud platforms promise to create value by preintegrating modular, industry-relevant solutions. Maximizing value from cloud investments will require automated operational scaling, access to cloud-native platform tools and adequate governance.

Assess:

- Augmented FinOps
- Cloud development environments
- Cloud sustainability
- Cloud-native
- Cloud-out to edge

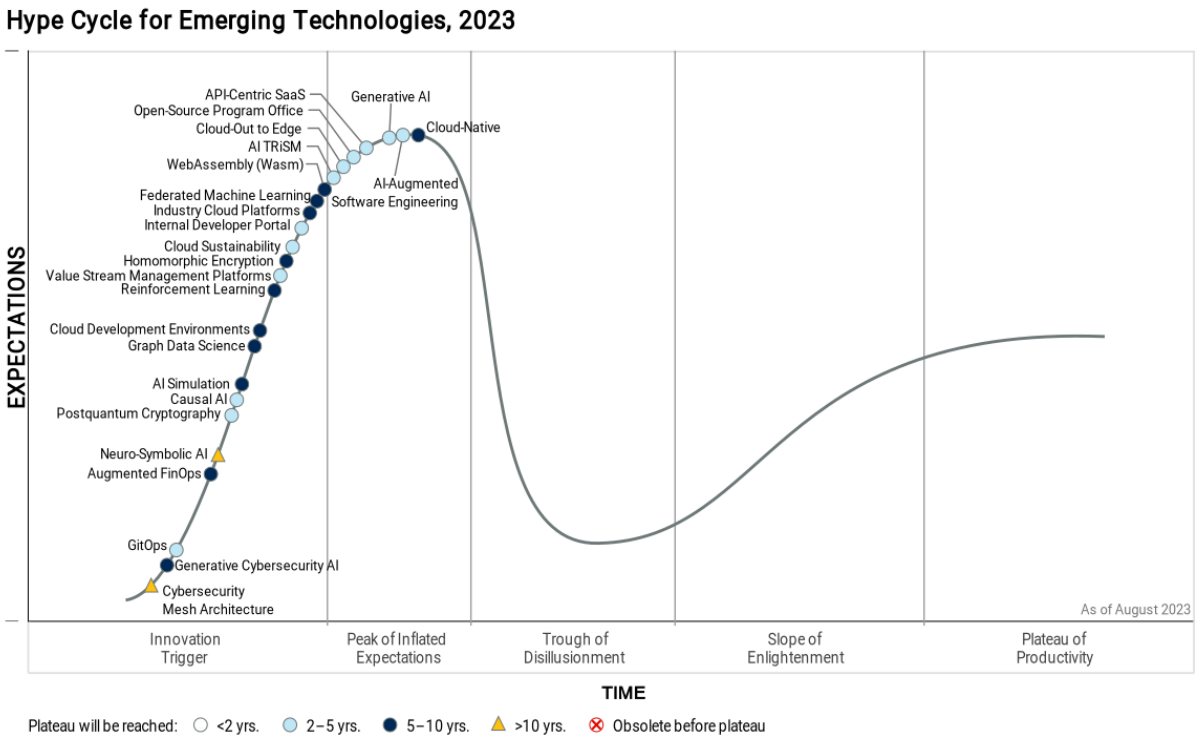
- Industry cloud platforms
- WebAssembly (Wasm)

Human-centric security and privacy: Humans remain the chief cause of security incidents and data breaches. Organizations can become resilient by implementing a human-centric security and privacy program. Such a program weaves a tight security and privacy fabric into the organization's digital design. The technologies highlighted below enable enterprises to create a culture of mutual trust and awareness of shared risks in decision making between many teams.

Explore:

- AI TRiSM
- Cybersecurity mesh architecture
- Generative cybersecurity AI
- Homomorphic encryption
- Postquantum cryptography

Figure 1: Hype Cycle for Emerging Technologies, 2023



The Priority Matrix

The Priority Matrix maps the benefit rating for each technology against the time it requires to achieve mainstream adoption. The benefit rating indicates the technology's potential, but the rating may not apply to all industries and organizations.

Most technologies have multiple use cases. To determine whether a technology will have a significant impact on your industry and organization, explore each use case. Prioritize technologies with the greatest potential benefit and prepare to launch a proof-of-concept project to demonstrate the feasibility of a technology for a specific use case.

When a technology can perform in a particular use case with reasonable quality, examine the other obstacles to deployment to determine when to deploy. Obstacles may be related to technical feasibility, organizational readiness and external factors (see Assessing Emerging Technology Adoption Readiness).

Examine technologies that offer more significant, near-term benefits because they can offer both strategic and tactical benefits. Explore technologies with longer-term benefits if they offer strategic value. Track technologies that are important to your organization by creating a technology radar (see Tool: How to Build an Emerging Technology Radar). Alternatively, use our Hype Cycle Builder tool to create a customized Hype Cycle for your organization (see Tool: Create Your Own Hype Cycle With Gartner's Hype Cycle Builder).

Emerging technologies are disruptive by nature, but the competitive advantage they provide isn't yet well-known or proven. Most will take more than five years, and some more than 10 years, to reach the Plateau of Productivity. But some technologies will mature in the near term, so you must understand the opportunities they present.

Table 1: Priority Matrix for Emerging Technologies, 2023

(Enlarged table in Appendix)

Benefit ↓	Years to Mainstream Adoption: Less Than 2 Years ↓	2 - 5 Years ↓	5 - 10 Years ↓	More Than 10 Years ↓
Transformational		AI-Augmented Software Engineering Generative AI	Augmented FinOps Generative Cybersecurity AI Homomorphic Encryption Industry Cloud Platforms WebAssembly (Wasm)	Cybersecurity Mesh Architecture
High		AI TRiSM API-Centric SaaS Causal AI Cloud-Out to Edge Cloud Sustainability GitOps Internal Developer Portal Open-Source Program Office Postquantum Cryptography Value Stream Management Platforms	AI Simulation Cloud Development Environments Cloud-Native Federated Machine Learning Graph Data Science Reinforcement Learning	Neuro-Symbolic AI
Moderate				
Low				

Source: Gartner (August 2023)

Off the Hype Cycle

The Hype Cycle for Emerging Technologies is not a typical Gartner Hype Cycle. It draws from an extremely broad spectrum of topics, and we intend it to be dynamic. It features many technologies for only a year or two, after which it stops tracking them to make room for other emerging technologies. Most of the technologies that we remove from this Hype Cycle continue to be tracked on other Hype Cycles. Refer to Gartner's broader collection of Hype Cycles for items of ongoing interest.

The technologies we've removed from this Hype Cycle in 2023 appear in the Hype Cycles indicated (information is correct at the time of publication):

- Autonomic systems — Hype Cycle for Artificial Intelligence, 2023; Hype Cycle for Enterprise Architecture, 2023; Hype Cycle for Monitoring and Observability, 2023; Hype Cycle for the Future of Enterprise Applications, 2023
- Cloud data ecosystems — Renamed “data ecosystems” and appears in Hype Cycle for Cloud Platform Services, 2023; Hype Cycle for Data and Analytics Programs and Practices, 2023; Hype Cycle for Data Management, 2023
- Computational storage — Hype Cycle for Storage and Data Protection Technologies, 2023
- Data observability — Hype Cycle for Data and Analytics Governance, 2023; Hype Cycle for Data and Analytics Programs and Practices, 2023; Hype Cycle for Data Management, 2023; Hype Cycle for Emerging Technologies in Finance, 2023; Hype Cycle for Finance Data and Analytics Governance, 2023; Hype Cycle for Midsize Enterprises, 2023; Hype Cycle for Monitoring and Observability, 2023
- Decentralized identity — Hype Cycle for Blockchain and Web3, 2023; Hype Cycle for Digital Identity, 2023; Hype Cycle for Emerging Technologies in Banking, 2023; Hype Cycle for Human Services in Government, 2023; Hype Cycle for Privacy, 2023
- Digital humans — At the time of publication, this innovation does not appear on any other Hype Cycle.
- Digital twin of a customer — Hype Cycle for Communications Service Provider Operations, 2023; Hype Cycle for Emerging Technologies in Banking, 2023; Hype Cycle for Emerging Technologies in Finance, 2023; Hype Cycle for Revenue and Sales Technology, 2023
- Dynamic risk governance — Hype Cycle for Cyber Risk Management, 2023; Hype Cycle for Legal and Compliance Technologies, 2023
- Foundation models — Hype Cycle for Artificial Intelligence, 2023; Hype Cycle for Data Science and Machine Learning, 2023; Hype Cycle for Emerging Technologies in Finance, 2023; Hype Cycle for Natural Language Technologies, 2023
- Generative design AI — Hype Cycle for Digital Grid Transformation Technologies, 2023; Hype Cycle for User Experience, 2023
- Internal talent marketplaces — Hype Cycle for Digital Workplace Applications, 2023; Hype Cycle for Emerging Technologies in Finance, 2023; Hype Cycle for HR Technology, 2023; Hype Cycle for Hybrid Work, 2023; Hype Cycle for Talent Acquisition Technologies, 2023; Hype Cycle for Workforce Transformation, 2023

- Machine learning code generation — Renamed “AI coding assistants” and appears in Hype Cycle for Software Engineering, 2023
- Metaverse — Hype Cycle for Blockchain and Web3, 2023; Hype Cycle for Customer Service and Support Technologies, 2023; Hype Cycle for Digital Government Services, 2023; Hype Cycle for Life Science Commercial Operations, 2023; Hype Cycle for Revenue and Sales Technology, 2023; Hype Cycle for Unified Communications and Collaboration, 2023
- Minimum viable architecture — Hype Cycle for Enterprise Architecture, 2023
- NFT — Hype Cycle for Blockchain and Web3, 2023
- Observability-driven development — Hype Cycle for Agile and DevOps, 2023; Hype Cycle for Monitoring and Observability, 2023; Hype Cycle for Site Reliability Engineering, 2023; Hype Cycle for Software Engineering, 2023
- OpenTelemetry — Hype Cycle for Container Technology, 2023; Hype Cycle for Monitoring and Observability, 2023
- Platform engineering — Hype Cycle for Agile and DevOps, 2023; Hype Cycle for Cloud Computing, 2023; Hype Cycle for Cloud Platform Services, 2023; Hype Cycle for Container Technology, 2023; Hype Cycle for Infrastructure Platforms, 2023; Hype Cycle for I&O Automation, 2023; Hype Cycle for Open-Source Software, 2023; Hype Cycle for Operating Models, 2023; Hype Cycle for Software Engineering, 2023
- Superapps — Hype Cycle for Application Architecture and Integration, 2023; Hype Cycle for Digital Government Services, 2023; Hype Cycle for Digital Workplace Applications, 2023; Hype Cycle for Hybrid Work, 2023; Hype Cycle for the Future of Enterprise Applications, 2023
- Web3 — Hype Cycle for Blockchain and Web3, 2023; Hype Cycle for Open-Source Software, 2023

On the Rise

Cybersecurity Mesh Architecture

Analysis By: Pete Shoard, Patrick Hevesi

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Cybersecurity mesh architecture (CSMA) is an emerging approach for architecting composable, distributed security controls that improve overall security effectiveness. It offers an approach to enabling secure, centralized security operations and oversight that emphasizes composable, independent security monitoring, predictive analytics and proactive enforcement, centralized intelligence and governance, and a common identity fabric.

Why This Is Important

CSMA aims to address the growing complexity of managing security tools, intelligence and identity solutions. Organizations must begin evolving toward a radically more flexible security architecture to prevent the impact of fast-emerging, evolving and retiring security tool categories and attack types. Planning to invest in composable, compatible and extensible security toolsets is essential to reduce cost and increase consistency.

Business Impact

CSMA offers a potential solution to problems currently suffered by defense-in-depth security architectures that most organizations employ. These are often made up of multiple point solutions that are poorly interconnected. CSMA addresses many challenges, including centralized exposure and security posture management, threat awareness, coordinated detection methodology and use cases, harmonized threat reporting and proactive response, and an increase in the efficiency of cross-tool collaboration.

Drivers

- Organizations increasingly require a broader perspective on the impact and likelihood of a threat or an exposure to a threat; it is this detail that is crucial in making better probusiness security decisions.

- IT security organizations can be overwhelmed when trying to stay ahead of new and more complex attacks, and when deploying the latest security tools to ever-expanding infrastructure. Teams are not able to implement the analytical capability required to be proactive and dynamic regarding their security enforcement and response decisions.
- Furthermore, these decisions are rarely fast enough to meet business needs.
- Effective security and identity management requires a layered and collaborative approach, but today's solutions are instead siloes that operate with insufficient knowledge of other tools and leave gaps. These silos are time-consuming to operate and monitor.
- Organizations understand and acknowledge the skills gaps and challenges in volumes of work, but do not have clear solutions to deal with these issues.
- Organizations are frustrated by the lack of integration and consistent visibility within their current security workbenches. Security and risk management leaders require an architecture that not only reacts to the current security issues (those that are visible in the organization), but provides a coordinated and holistic approach to complex security problems.
- Creating a collaborative ecosystem of security tools will address inconsistency and help understand and minimize the exposure that is consistent with business expectations.

Obstacles

- As CSMA emerges and vendors add support for the architecture principles to their products, vendor lock-in will likely be a concern. If a proprietary approach is employed, it may serve to block, rather than facilitate, cross-tool integration; then gaps in coverage will likely appear, and this inflexibility will drive up cost.
- Those organizations that choose to create their own CSMA construct will likely need significant engineering effort to integrate disparate products and may suffer if the security industry moves toward a set of standards for interoperability after significant custom integration work has been completed.
- Currently, there are no vendors that offer what might be described as a CSMA solution. Features and requirements of the reference architecture continue to evolve in response to consumer IT advancement and security technology consolidation as a result of vendor acquisitions and partnerships.

User Recommendations

- Position your organization for a future of rapid change. Add purchasing requirements that focus on integration and interoperability of multivendor tools.
- Mature your security infrastructure by selecting point product vendors that are aligning to the CSMA reference architecture, have fully developed advanced APIs, complete adherence to modern security standards and have integrations into security partner networks.
- Evolve your identity infrastructure to an identity fabric by removing silos to achieve dynamic real-time identity capabilities that incorporate a more complete set of context and risk signals (such as device proximity, posture, biometrics and location).
- Improve your responsiveness by centralizing your policy, posture and playbook management along with building an integrated “single starting pane of glass” view into your CSMA.

Gartner Recommended Reading

The Future of Security Architecture: Cybersecurity Mesh Architecture (CSMA)

How to Start Building a Cybersecurity Mesh Architecture

2023 Planning Guide for Security

Emerging Technology Horizon for Information Security, 2022

Generative Cybersecurity AI

Analysis By: Jeremy D'Hoinne, Avivah Litan, Mark Horvath, Wilco van Ginkel

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Generative cybersecurity AI technologies generate new derived versions of security-related and other relevant content, strategies, designs and methods by learning from large repositories of original source data. Generative cybersecurity AI can be delivered as a public or privately hosted cloud service or embedded with security management interfaces. It can also integrate with software agents to take action.

Why This Is Important

Enterprises witness many applications leveraging foundation models that can read multimodal objects (such as sensory data and images), following the first applications based on large language models (LLMs).

Cybersecurity technology providers can exploit generative cybersecurity AI to improve existing workflows, be a proxy of existing analytics, and generate security configuration or realistic attack data. Soon, applications will include autonomous agents, which can work using high-level guidance without a need for frequent prompting.

Business Impact

Existing vendors and new startups will add generative cybersecurity AI, expanding or replacing features. They will start implementing it with resource-intensive tasks, such as incident response, exposure or risk management, or code analysis.

Organizations will benefit from generative cybersecurity AI as it can improve efficiency and shorten response times to cybersecurity risks and threats. The pace of adoption will vary across industries and geographies due to security and privacy concerns.

Drivers

- ChatGPT is one of the most hyped and fastest-adopted AI technologies ever. It relies on generative AI foundation models, which are largely trained on massive internet datasets.
- Security operations center (SOC) teams cannot keep up with the deluge of security alerts they must constantly review, and are missing key threat indicators in the data.
- Risk analysts need to speed up risk assessments, and be more agile and adaptable through increased automation and prepopulation of risk data in context.
- Organizations continue to experience skill shortages and look for opportunities to automate resource-intensive cybersecurity tasks. Use cases for the application of generative AI include: synthesizing and analyzing threat intelligence; generating remediation suggestions for application security, cloud misconfigurations and configuration changes to adjust to threats; generating scripts and codes generation; implementing secure code agents; identifying and graphing key security events in logging systems; conducting risk and compliance identification and analysis; automating the first steps in incident response; tuning of security configuration adjustment; creating general best practice guidance.
- Generative cybersecurity AI augments existing continuous threat exposure management (CTEM) programs by better aggregating, analyzing and prioritizing inputs. It also generates realistic scenarios for validation.
- Generative AI offerings include the ability to fine-tune models, develop applications using prompt engineering and integrate with prepackaged tools and plugins through APIs. These possibilities open up a path for providers to add generative cybersecurity AI.
- Microsoft has already demonstrated a preview version of its security co-pilot feature, which is expected to drive competitors to embed similar approaches.
- Security program performance solutions and activities can solve their increasing demand for business alignment. Further, they can perform scenario planning for budget (re)allocation, and efficiency and effectiveness indicators and corrections.

Obstacles

- The cybersecurity industry is already plagued with false positives. Early examples of “hallucinations” and inaccurate responses will cause organizations to be cautious about adoption or limit the scope of their usage.
- Best practices and tooling to implement responsible AI, privacy, trust, security and safety for generative AI applications do not fully exist yet.
- Privacy and intellectual property concerns could prevent sharing and usage of business- and threat-related data, reducing the accuracy of generative cybersecurity AI outputs.
- As generative AI is still emerging, establishing the trust required for its wider adoption will take time. This is especially true for the skill augmentation use cases, as you would need the skills you are supposed to augment, in order to ensure the recommendations are good.
- Uncertainty on laws and regulations related to generative AI may slow down adoption in some industries, for example regulated industries in EU countries subject to GDPR compliance.

User Recommendations

- Pick initial use cases carefully. First implementations might have a higher error rate than more mature techniques already in place.
- Monitor the addition of generative AI features from your existing providers and beware of “generative AI washing.” Don’t pay a premium before obtaining measurable results.
- Choose fine-tuned models that align with the relevant security use case or fine-tune in-house models from base models offered by the providers.
- Refrain from sharing sensitive and confidential data with hosted models until verifiable privacy assurances are provided by the host.
- Apply AI security frameworks, such as AI TRiSM. Work with your legal team on data privacy and copyright issues.
- Implement a documented approval workflow for allowing new generative cybersecurity AI experiments.
- Make it mandatory from a policy standpoint that any content (that is, configuration or code) generated by an AI is fully documented, peer-reviewed by humans and tested before it is implemented. If not possible, consider any AI-generated content as “Draft Only” when used for critical use cases.

Gartner Recommended Reading

4 Ways Generative AI Will Impact CISOs and Their Teams

Innovation Insight for Generative AI

Market Guide for AI Trust, Risk and Security Management

GitOps

Analysis By: Paul Delory, Arun Chandrasekaran

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

GitOps is a type of closed-loop control system for cloud-native applications. The term is often used more expansively, usually as a shorthand for automated operations or CI/CD, but this is incorrect. According to the canonical OpenGitOps standard, the state of any system managed by GitOps must be: (1) expressed declaratively, (2) versioned and immutable, (3) pulled automatically, and (4) continuously reconciled. These ideas are not new, but new tools and practices now bring GitOps within reach.

Why This Is Important

GitOps can be transformative. GitOps workflows deploy a verified and traceable configuration (such as a container definition) into a runtime environment, bringing code to production with only a Git pull request. All changes flow through Git, where they are version-controlled, immutable and auditable. Developers interact only with Git, using abstract, declarative logic. GitOps extends a common control plane across Kubernetes (K8s) clusters, which is increasingly important as clusters proliferate.

Business Impact

By operationalizing infrastructure as code, GitOps enhances availability and resilience of services:

- GitOps can be used to improve version control, automation, consistency, collaboration and compliance.
- Artifacts are reusable and can be modularized.
- Configuration of clusters or systems can be updated dynamically. All of this translates to business agility and a faster time to market.
- GitOps artifacts are version-controlled and stored in a central repository, making them easy to verify and audit.

Drivers

- **Kubernetes adoption and maturity:** GitOps must be underpinned by an ecosystem of technologies, including tools for automation, infrastructure as code, continuous integration/continuous deployment (CI/CD), observability and compliance. Kubernetes has emerged as a common substrate for cloud-native applications. This provides a ready-made foundation for GitOps. As Kubernetes adoption grows within the enterprise, so can GitOps, too.
- **Need for increased speed and agility:** Speed and agility of software delivery are critical metrics that CIOs care about. As a result, IT organizations are pursuing better collaboration between infrastructure and operations (I&O) and development teams to drive shorter development cycles, faster delivery and increased deployment frequency. This will enable organizations to respond immediately to market changes, handle workload failures better, and tap into new market opportunities. GitOps is the latest way to drive this type of cross-team collaboration.
- **Need for increased reliability:** Speed without reliability is useless. The key to increased software quality is effective governance, accountability, collaboration and automation. GitOps can enable this through transparent processes and common workflows across development and I&O teams. Automated change management helps to avoid costly human errors that can result in poor software quality and downtime.
- **Talent retention:** Organizations adopting GitOps have an opportunity to upskill existing staff for more automation- and code-oriented I&O roles. This opens up opportunities for staff to learn new skills and technologies, resulting in higher employee satisfaction and retention.
- **Cultural change:** By breaking down organizational silos, development and operations leaders can build cross-functional knowledge and collaboration skills across their teams to enable them to work effectively across boundaries.
- **Cost reduction:** Automation of infrastructure eliminates manual tasks and rework, improving productivity and reducing downtimes, both of which can contribute to cost reduction.

Obstacles

- **Prerequisites:** GitOps is only for cloud-native applications. Many GitOps tools and techniques assume the system is built on Kubernetes (frequently, they also assume that a host of other technologies are built on top of K8s). By definition, GitOps requires software agents to act as listeners for changes and help to implement them. GitOps is possible outside Kubernetes, but in practice K8s will almost certainly be used. Thus, GitOps is necessarily limited in scope.
- **Cultural change:** GitOps requires a cultural change that organizations need to invest in. IT leaders need to embrace process change. This requires discipline and commitment from all participants to doing things in a new way.
- **Skills gaps:** GitOps requires automation and software development skills, which many I&O teams lack.
- **Organizational inertia:** GitOps requires collaboration among different teams. This requires trust among these teams for GitOps to be successful.

User Recommendations

- **Target cloud-native workloads initially:** Your first use case for GitOps should be operating a containerized, cloud-native application that is already using both Kubernetes and a continuous delivery platform such as Flux or ArgoCD.
- **Build an internal operating platform:** This is the foundation of your GitOps efforts. Your platform should manage the underlying infrastructure and deployment pipelines, while enforcing security and policy compliance.
- **Embed security into GitOps workflows:** Security teams need to shift left, so the organization can build holistic CI/CD pipelines that deliver software and configure infrastructure, with security embedded in every layer.
- **Be wary of vendors trying to sell you GitOps:** GitOps isn't a product you can buy, but a workflow and a mindset shift that becomes part of your overall DevOps culture. Tools that expressly enable GitOps can be helpful; but GitOps can be done with nothing more than standard continuous delivery tools that support Git-based automation.

Sample Vendors

GitLab; Harness; Red Hat; Upbound; Weaveworks

Gartner Recommended Reading

Innovation Insight: Top 4 Use Cases for GitOps

Is Using GitOps-Style Automation With Kubernetes Right for Me?

How to Scale DevOps Workflows in Multicluster Kubernetes Environments

Designing and Operating DevOps Workflows to Deploy Containerized Applications With Kubernetes

Automate the Application Delivery Value Stream

Augmented FinOps

Analysis By: Adam Ronthal, Dennis Smith

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

FinOps applies the traditional DevOps concepts of agility, continuous integration and deployment, and end-user feedback to financial governance, budgeting and cost optimization efforts. Augmented FinOps automates this process through the application of artificial intelligence (AI) and machine learning (ML) practices — predominantly in the cloud — to enable environments that automatically optimize cost based on defined business objectives expressed in natural language.

Why This Is Important

In the cloud, it is now possible to assess the cost of a specific workload or collection of workloads assigned to a project. However, price/performance — the primary measure of cloud efficiency — is difficult to assess due to the complexity and diversity of choice in underlying cloud infrastructure and service offerings and a lack of consistency in pricing models. Augmented FinOps can automate this process by applying AI/ML techniques.

Business Impact

The automation of cloud budget planning and financial operations will allow businesses to express their objectives — ideally in natural language — and allow their cloud ecosystems to automatically optimize the underlying cloud resources to meet those objectives. This will result in more efficient use of resources and, therefore, optimal spend by reducing/eliminating misaligned or poor use of cloud infrastructure and service offerings.

Drivers

- Practitioners are increasingly realizing that cloud is fundamentally a complex cost optimization exercise.
- Cloud adopters have a strong desire for transparency into cloud spending.
- Buyer inexperience is leading to either under-provisioning and associated resource contention or overprovisioning and spending more than is needed.
- Vendors are positioning cost-effectiveness as a competitive differentiator in their go-to-market strategies.
- Practitioners need to reduce the unpredictability of cloud spending when using cloud infrastructure and services for analytics, operational database management systems (DBMSs), data lakes and other applications, including custom IT infrastructure.
- Consumption-based usage remains common in earlier stages of cloud adoption, driving the need for augmented FinOps, although commit-based usage mitigates some unpredictability.
- Cost overruns are often obscured, downplayed, or dismissed by line of business implementers, requiring augmentation to achieve holistic and comprehensive cost optimization.
- Automation of financial governance controls in cloud environments provides increased predictability and cost optimization with less operational effort.
- Solid financial governance frameworks are positioning organizations to take advantage of FinOps.
- Emergence of specific roles — like FinOps practitioner or cloud economist — focused on FinOps practices and cost optimization means organizations have the expertise to address augmented FinOps.
- Owing to their complexity, cloud environments are ideally suited for the application of ML and AI methods to automate processes and track price and performance.
- Core FinOps capabilities are being delivered in three ways: Homegrown solutions, cloud service provider (CSP) instrumentation and third-party vendors. Increasingly practitioners are seeking out third-party or CSP tools to address their needs. All of these have a broad objective of adopting augmented capabilities as a means of competitive differentiation.

Obstacles

- Cloud service provider pricing models remain needlessly complex and diverse.
- Cloud ecosystems are (and will remain) open to third-party participants, which implies multiple commercial arrangements with multiple providers.
- Standards for cloud cost, usage and billing data like the FinOps Foundation's FOCUS proposal have yet to be broadly adopted. APIs for communicating performance data within the context of a broader ecosystem have yet to emerge. Both of these are required to assess the primary measure of success: price/performance.

User Recommendations

- Seek out service offerings to automate (via AI/ML) performance, consumption and pricing options. Increasingly, incorporate these capabilities into cloud data ecosystems that will learn from consumption patterns as they seek to optimize the underlying resources, and by extension, cloud spending through orchestration and optimization.
- Apply Gartner's FinOps Maturity Model to assess FinOps offerings in terms of their ability to address the following core capabilities: Observe, report, recommend, predict and optimize. The last three introduce augmented FinOps capabilities.
- Plan to use multiple tools to address the full scope of requirements. Many tools are broad in reach, but do not go deep into prescriptive recommendations. Others are tightly scoped and provide very targeted optimizations. Expect to spend time combining multiple tools to achieve broad and deep capabilities.

Sample Vendors

Acceldata; Anodot; Apptio; Capital One Software; Densify; Enteros; Finout; OtterTune; Sync Computing; Unravel Data

Gartner Recommended Reading

How to Identify Solutions for Managing Costs in Public Cloud IaaS

A Guidance Framework for Selecting Cloud Management Tools

Emerging Tech: Data Management Product Leaders Must Implement Augmented FinOps in Their Cloud Solutions

CDAOs and CFOs Must Drive Business Value in the Cloud Through Collaboration

Financial Governance Is Essential to Successful Cloud Data and Analytics

Neuro-Symbolic AI

Analysis By: Erick Brethenoux, Afraz Jaffri

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Neuro-symbolic AI is a form of composite AI that combines machine learning methods and symbolic systems (for example, knowledge graphs) to create more robust and trustworthy AI models. This combination allows statistical patterns to be combined with explicitly defined rules and knowledge to give AI systems the ability to better represent, reason and generalize concepts. This approach provides a reasoning infrastructure for solving a wider range of business problems more effectively.

Why This Is Important

Neuro-symbolic AI is important because it addresses limitations in current AI systems, such as incorrect output (hallucinations in large language models [LLMs]), generalization to a variety of tasks and explaining the steps that led to an output. This leads to more powerful, versatile and interpretable AI solutions and allows AI systems to tackle more complex tasks with humanlike reasoning.

Business Impact

Neuro-symbolic AI will have an impact on the efficiency, adaptability and reliability of AI systems used across business processes. The integration of logic and multiple reasoning mechanisms brings down the need for ever larger ML models and their supporting infrastructure. These systems will rely less on the processing of huge amounts of data, making AI agile and resilient. Decision making can be augmented and automated using neuro-symbolic approaches with less risk of unintended consequences.

Drivers

- Limitations of AI models that rely purely on machine learning techniques that focus on correlation over understanding and reasoning. The newest generation of LLMs is well-known for its tendency to give factually incorrect answers or produce unexpected results.
- The need for explanation and interpretability of AI outputs that are especially important in the regulated industry use cases and in systems that use private data.
- The need to move toward semantics over syntax in systems that deal with real-world entities in order to ground meaning to words and terms in specific domains.
- The set of tools available to combine different types of AI models is increasing and becoming easier to use for developers, data scientists and end users. The dominant approach is to chain together results from different models (composite AI) rather than using single models that are neuro-symbolic in nature.
- The integration of multiple reasoning mechanisms necessary to provide agile AI systems eventually leading to adaptive AI systems.

Obstacles

- Most neuro-symbolic AI methods and techniques are being developed in academia or industry research labs. Despite the increase in tools available, there are still limited implementations in business or enterprise settings.
- There are no agreed-upon techniques for implementing neuro-symbolic AI and disagreements continue between researchers and practitioners on the effectiveness of combining approaches; despite the emergence of real-world use cases.
- The commercial and investment trajectory for AI startups allocates almost all capital to deep learning approaches leaving only those willing to bet on the future to invest in neuro-symbolic AI development.
- Popular media and academic conferences do not give as much exposure to the neuro-symbolic AI movement as compared to other approaches, for now.

User Recommendations

- Adopt composite AI approaches when building AI systems by utilizing a range of techniques that increase the robustness and reliability of AI models. Neuro-symbolic AI approaches will fit into a composite AI architecture.
- Dedicate time to learning and applying neuro-symbolic AI approaches by identifying use cases that can benefit from these approaches.
- Invest in data architecture that can leverage the building blocks for neuro-symbolic AI techniques such as knowledge graphs and agent-based techniques.

Sample Vendors

Google Deepmind; Elemental Cognition; IBM; Microsoft; RelationalAI; Wolfram|Alpha

Gartner Recommended Reading

Innovation Insight: AI Simulation

Innovation Insight for Composite AI

Predicts 2023: Simulation Combined With Advanced AI Techniques Will Drive Future AI Investments

AI Simulation

Analysis By: Leinar Ramos, Anthony Mullen, Pieter den Hamer, Jim Hare

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

AI simulation is the combined application of AI and simulation technologies to jointly develop AI agents and the simulated environments in which they can be trained, tested and sometimes deployed. It includes both the use of AI to make simulations more efficient and useful, and the use of a wide range of simulation models to develop more versatile and adaptive AI systems.

Why This Is Important

Increased complexity in decision making is driving demand for both AI and simulation. However, current AI faces challenges, as it is brittle to change and requires a lot of data. Conversely, realistic simulations can be expensive and difficult to build and run. To resolve these challenges, a growing approach is to combine AI and simulation: Simulation is used to make AI more robust and compensate for a lack of training data, and AI is used to make simulations more efficient and realistic.

Business Impact

AI simulation can bring:

- Increased AI value by broadening its use to cases where data is scarce, using simulation to generate synthetic data (for example, robotics and self-driving cars)
- Greater efficiency by leveraging AI to decrease the time and cost to create and use complex and realistic simulations
- Greater robustness and adaptability by using simulation to generate diverse scenarios to increase AI performance in uncertain environments
- Decreased technical debt by reusing simulation environments to train future AI models

Drivers

- **Limited availability of AI training data is increasing the need for synthetic data techniques, such as simulation.** Simulation is uniquely positioned among synthetic data alternatives in its ability to generate diverse datasets that are not constrained by a fixed "seed" dataset to generate synthetic data from.
- **Advances in capabilities are making simulation increasingly useful for AI.** Simulation capabilities have been rapidly improving, driven both by increased computing performance and more efficient techniques. This has made simulation environments a key part of the training pipeline of some of the most advanced real-world AI use cases, such as robotics and self-driving cars.
- **The growing complexity of decision making is increasing the interest in AI simulation.** Simulation is able to generate diverse "corner case" scenarios that do not appear frequently in real-world data, but that are still crucial to train and test AI to perform well on uncertain environments. As the complexity of the environments and decision making goes up, the ability to build AI systems that are robust becomes more important.
- **Increased technical debt in AI is driving the need for the reusable environments that simulation provides.** Current AI focuses on building short-lived AI models with limited reuse, accumulating technical debt. Organizations will increasingly deploy hundreds of AI models, which requires a shift in focus toward building persistent, reusable environments where many AI models can be trained, customized and validated. Simulation environments are ideal since they are reusable, scalable, and enable the training of many AI models at once.
- **The growing sophistication of simulation drives the use of AI to make it more efficient.** Modern simulations are resource-intensive. This is driving the use of AI to accelerate simulation, typically by employing AI models that can replace parts of the simulation without running resource-intensive step-by-step numerical computations.

Obstacles

- **Gap between simulation and reality:** Simulations can only emulate — not fully replicate — real-world systems. This gap will reduce as simulation capabilities improve, but it will remain a key factor. Given this gap, AI models trained in simulation might not have the same performance once they are deployed: differences in the simulation training dataset versus real-world data can impact models' accuracy.
- **Complexity of AI simulation pipelines:** The combination of AI and simulation techniques can result in more complex pipelines that are harder to test, validate, maintain and troubleshoot.
- **Limited readiness to adopt AI simulation:** A lack of awareness among AI practitioners about leveraging simulation capabilities can prevent organizations from implementing an AI simulation approach.
- **Fragmented vendor market:** The AI and simulation markets are fragmented, with few vendors offering combined AI simulation solutions, potentially slowing down the deployment of this capability.

User Recommendations

- Complement AI with simulation to optimize business decision making or to overcome a lack of real-world data by offering a simulated environment for synthetic data generation or reinforcement learning.
- Complement simulation with AI by applying deep learning to accelerate simulation and generative AI to augment simulation.
- Create synergies between AI and simulation teams, projects and solutions to enable a next generation of more adaptive solutions for ever-more complex use cases. Incrementally build a common foundation of more generalized and complementary models that are reused across different use cases, business circumstances and ecosystems.
- Prepare for the combined use of AI, simulation and other relevant techniques, such as graphs, natural language processing or geospatial analytics, by prioritizing vendors that offer platforms that integrate different AI techniques (composite AI), as well as simulation.

Sample Vendors

Altair; Ansys; Cosmo Tech; Epic Games; MathWorks; Microsoft; NVIDIA; Rockwell Automation; The AnyLogic Company; Unity

Gartner Recommended Reading

Innovation Insight: AI Simulation

Predicts 2023: Simulation Combined With Advanced AI Techniques Will Drive Future AI Investments

Cool Vendors in Simulation for AI

Causal AI

Analysis By: Pieter den Hamer, Leinar Ramos, Ben Yan

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Causal artificial intelligence (AI) identifies and utilizes cause-and-effect relationships to go beyond correlation-based predictive models and toward AI systems that can prescribe actions more effectively and act more autonomously. It includes different techniques, such as causal graphs and simulation, that help uncover causal relationships to improve decision making.

Why This Is Important

AI's ultimate value comes from helping people take better actions. Machine learning (ML) makes predictions based on statistical relationships (correlations), regardless of whether these are causal. This approach is fine for prediction, but predicting an outcome is not the same as understanding what causes it and how to improve it. Causal AI is crucial when we need to be more prescriptive to determine the best actions to influence specific outcomes. Causal AI techniques help make AI more autonomous, explainable, robust and efficient.

Business Impact

Causal AI leads to:

- Greater decision augmentation and autonomy in AI systems by estimating intervention effects
- Greater efficiency by adding domain knowledge to bootstrap AI models with smaller datasets
- Better explainability by capturing easy-to-interpret cause-and-effect relationships
- More robustness and adaptabilty by leveraging causal relationships that remain valid in changing environments
- The ability to extract causal knowledge with less costly and time-consuming experiments
- Reduced bias in AI systems by making causal links more explicit

Drivers

- **Analytics demand is shifting from predictive (what is likely to happen) to more prescriptive (what should be done) capabilities.** Making accurate predictions will remain key, but a causal understanding of how to affect predicted outcomes will be increasingly important.
- **AI systems increasingly need to act autonomously to generate business value,** particularly for time-sensitive and complex use cases, where human intervention is not feasible. This autonomy will only be possible by AI understanding what impact actions will have and how to make effective interventions.
- **Limited data availability for certain use cases is pushing organizations toward more data-efficient techniques like causal AI.** Causal AI leverages human domain knowledge of cause-and-effect relationships to bootstrap AI models in small-data situations.
- **The growing complexity of use cases and environments where AI is applied requires more robust AI techniques.** Causal structure changes much more slowly than statistical correlations, making causal AI more robust and adaptable in fast-changing environments. The volatility of the last few years has exposed the brittleness of correlation-based AI models across industries. These models have struggled to adapt because they were trained under a very different context.
- **The need for greater AI trust and explainability is driving interest in models that are more intuitive to humans.** Causal AI techniques, such as causal graphs, make it possible to be explicit about causes and explain models in terms that humans understand.
- **The next step in AI requires causal AI.** Current deep learning models and, in particular, generative AI have limitations in terms of their reliability and ability to reason. A composite AI approach that complements generative AI with causal AI — in particular, causal knowledge graphs — offers a promising avenue to bring AI to a higher level.

Obstacles

- **Causality is not trivial.** Not every phenomenon is easy to model in terms of its causes and effects. Causality might be unknown, regardless of AI use.
- **The quality of a causal AI model depends on its causal assumptions and on the data used to build it.** This data is susceptible to bias and imbalance. Just because a model is causal doesn't mean that it will outperform correlation-based ones.
- **Causal AI requires technical and domain expertise to properly estimate causal effects.** Building causal AI models is often more difficult than building correlation-based predictive models, requiring active collaboration between domain experts and AI experts.
- **AI experts might be unaware of causality methods.** If AI experts are overly reliant on data-driven models like ML, organizations could get pushback when looking to implement causal AI.
- **The vendor landscape is nascent, and enterprise adoption is currently low.** Clearly, this represents a challenge when organizations are running initial causal AI pilots and identifying specific use cases where causal AI is most relevant.

User Recommendations

- Acknowledge the limitations of correlation-based AI and ML approaches, which focus on leveraging correlations and mostly ignore causality. These limitations also apply to most generative AI techniques, including foundation models such as GPT-4.
- Use causal AI when you require more augmentation and automation in decision intelligence — i.e., when AI is needed not only to generate predictions, but also to understand how to affect the predicted outcomes. Examples include customer retention programs, marketing campaign allocation and financial portfolio optimization, as well as smart robotics and autonomous systems.
- Select different causal AI techniques depending on the complexity of the specific use case. These include causal rules, causal graphs and Bayesian networks, simulation, and ML for causal learning.
- Educate your data science teams on causal AI. Explain the difference between causal and correlation-based AI, and cover the range of techniques available to incorporate causality.

Sample Vendors

Actable AI; causaLens; Causality Link; CML Insight; Geminos Software; IBM; Lucid.AI; Qualcomm; SCALNYX; Xplain Data

Gartner Recommended Reading

Innovation Insight: Causal AI

Innovation Insight for Composite AI

Innovation Insight for Decision Intelligence Platforms

Building a Digital Future: Autonomic Business Operations

Case Study: Causal AI to Maximize the Efficiency of Business Investments (HDFC Bank)

Postquantum Cryptography

Analysis By: Mark Horvath, Matthew Brisse

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Definition:

Postquantum cryptography (PQC), also called quantum-safe cryptography, are algorithms designed to secure against both classical and quantum-computing attacks. PQC will replace existing asymmetric encryption, which will weaken over the next decade, deprecating existing classical encryption methodologies and processes.

Why This Is Important

PQC offers organizations a higher level of cryptographic protection, which will remain strong as quantum computers enter the mainstream.

Existing asymmetric algorithms like Diffie-Hellman, RSA and ECC are vulnerable and will be unsafe to use by the end of the current decade, requiring replacement for common cryptographic functions, such as digital signatures, public key encryption and key exchanges.

Business Impact

PQC has the following impacts:

- With the advent of stronger quantum computers, existing asymmetric algorithms must be replaced with quantum-safe ones. This includes all network, file and data encryption, IAM, as well as any other uses of asymmetric cryptography.
- There are no drop-in alternatives for existing cryptographic algorithms, leading to discovery, categorization and reimplementation efforts.
- As new algorithms have different performance characteristics, current applications must be retested and, in some cases, rewritten.

Drivers

- Existing asymmetric encryption algorithms will become vulnerable to quantum decryption attacks by the end of the decade, potentially requiring reencryption of all data where the risk of exposure of the symmetric keys or tokens is considered important.
- Governments around the world are preparing and issuing mandates and legal frameworks requiring government agencies and enterprises to start devising PQC strategies. For example, in the U.S., Quantum Computing Cybersecurity bill requires owners and operators of national security systems and organizations supplying to the U.S. government to start using postquantum algorithms.
- “Harvest now and decrypt later” attacks are an ongoing concern, leading to the urgency to implement PQC security measures sooner rather than later.
- Secondary uses of new encryption (e.g., homomorphic encryption, stateful signatures, etc.) will offer new business opportunities beyond data protection.
- Once PQC is adopted by an organization, data should be secure for the foreseeable future.

Obstacles

- Most organizations don't know how cryptography functions within their organization, where keys and algorithms are used, or how secrets are stored and managed. Swapping them out for new algorithms will be challenging.
- Encrypted file sizes and digital signatures for new algorithms are typically much bigger than existing equivalents, necessitating hardware and network infrastructure improvements.
- New PQC algorithms will require new standards. The current set of PQC candidates' standards are expected to be released in late 2023 or early 2024, while fresh algorithm development will continue for the rest of the decade, affecting hardware, firmware, software and credentials used along with supported algorithms.
- Most vendors are typically unprepared when it's time to upgrade the cryptography and often require some pushing from their clients to recognize the demand.
- Some very important protocols lack built-in crypto-agility. For instance, no one is developing plans on how to incorporate new algorithms into WS-Security, which is used to safeguard SOAP APIs (a crucial type of API for all financial transactions).

User Recommendations

- Build a cryptographic metadata database of all in-use cryptographic algorithms.
- Develop crypto policies for easing the transition to new algorithms.
- Perform an exercise for data identifying the expected end-of-life targets in the short, medium and long-term time scales, and create a key life cycle policy to reflect risks to asymmetric and symmetric crypto keys.
- Create a transition phase plan identifying which algorithms are suitable for particular use cases.
- Implement transitional crypto policies for when algorithms should be replaced and which new algorithms should be used in each use case.
- Implement crypto-agile application development and stage to production after extensive testing.
- Vet and test new PQC algorithms to understand their characteristics and uses.
- Implement crypto-agility initiatives with an object-based approach to address future changes in PQC algorithm updates and replacement.
- Prioritize business impact potential when selecting potential PQC use cases.

Sample Vendors

Amazon; Google; IBM; ISARA; Microsoft; Qrypt; Quantum Xchange; SandboxAQ

Gartner Recommended Reading

Preparing for the Quantum World With Crypto-Agility

Emerging Tech: How to Make Money From Quantum Computing

Emerging Tech: Critical Insights on Quantum Computing

Infographic: How Use Cases Are Developed and Executed on a Quantum Computer

Cloud Development Environments

Analysis By: Manjunath Bhat

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Cloud development environments (CDEs) provide remote, ready-to-use access to a cloud-hosted development environment with minimal effort for setup and configuration. This decoupling of the development workspace from the physical workstation enables a low-friction, consistent developer experience. CDEs comprise elements of a traditional IDE, such as code editing, debugging, code review and code collaboration. They increasingly include ML-based coding assistants and integrate with DevOps platforms.

Why This Is Important

CDEs provide consistent, secure developer access to preconfigured development workspaces. This frees them from setting up their own local environments, eliminating the need to install and maintain dependencies, software development kits, security patches and plug-ins. CDEs are prepackaged with language tooling for multiple programming languages, enabling teams to write code for different application stacks with standardized and templatized workflows.

Business Impact

CDEs are becoming popular due to their seamless integration with Git-based repositories and continuous integration/continuous delivery (CI/CD) tools, thus enhancing developer productivity and developer experience. They also enable security and platform teams to govern and secure user access to development environments, even from personally-owned devices. This mitigates the security and operational risks. CDEs provide a scalable way to support a distributed workforce and hybrid work environments.

Drivers

Gartner predicts that by 2026, 60% of cloud workloads will be built and deployed using CDEs. Four factors are driving their increased adoption:

- Remote work and remote onboarding of software developers create a need for a frictionless onboarding experience. The ability to share the development environment among distributed team members makes remote debugging and pair programming easier.

- Organizations increasingly need custom configured hardware to support specific application architectures, such as Apple M1/M2 silicon for mobile app development or GPU support for data and analytics workloads.
- The ability to centrally manage, govern and secure development environments has become especially important to minimize the threat of software supply chain attacks. In addition, CDEs make it easier to support and secure bring-your-own-device use cases.
- Automating DevOps workflows introduces more plug-ins, extensions and API integrations, which makes the distribution of updates and configurations to local machines cumbersome and unreliable.
- Building modern, distributed applications such as mesh services, event-driven apps and Kubernetes applications introduce complexities, making it harder to develop and test on local machines.

Obstacles

- CDEs incur costs in addition to what an organization may already be paying for DevOps tooling. The cost can be prohibitive for teams that rely on open-source development tools for application development and delivery on local machines.
- Connectivity presents another obstacle. Poor or inconsistent internet speeds adversely affect developer experience. In many cases, developers simply like to use their own local machines to get their work done. That allows them to use their own plugins, editors and scripts, all of which will be difficult through a browser-based interface.
- Security and compliance policies can prevent the use of cloud for development in some organizations. This can rule out CDEs that rely on public cloud services. Note that CDEs don't necessarily have to be provisioned in public cloud environments.
- There may be resistance from developers since it may impede their ability to research, experiment and innovate using full or elevated permissions on local development machines.

User Recommendations

Software engineering leaders should:

- Pilot the use of CDEs when their teams are developing with cloud-hosted source repositories. Plan for downtime resulting from service outages, since CDEs rely on the remote development environment being up and running.
- Ensure that CDEs are part of an overall developer self-service platform with centralized governance. The benefit of centralized governance and developer agility can be a win-win for platform and product teams.
- Use CDEs as one of the "quick wins" to improve the onboarding experience for new developers and reduce ramp-up time.
- Work with security experts and enforce strong authentication and authorization policies to mitigate security risks. CDEs present an additional, high-value attack vector, as they become the pipeline through which intellectual property flows.

Sample Vendors

Amazon Web Services; Codeanywhere; Coder; GitHub; Gitpod; Google; JetBrains Space; Red Hat; Replit; Strong Network

Gartner Recommended Reading

Quick Answer: How to Create a Frictionless Onboarding Experience for Software Engineers

Infographic: Platforms and Tools to Scale the Delivery of High-Quality Software

Innovation Insight for Internal Developer Portals

Innovation Insight for ML-Powered Coding Assistants

Graph Data Science

Analysis By: Afraz Jaffri, Svetlana Sicular

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Graph data science (GDS) is a discipline in which data science techniques are applied to graph data structures to identify behavioral characteristics that can be used to build predictive and prescriptive models. Graph data science and machine learning algorithms compute distances and paths, similarity, and communities; simulate the effects of changes in the graph; and allow predictions to be made for inferring new nodes or edges or classifying whole graph structures.

Why This Is Important

Graphs are being used in a wide variety of scenarios, such as financial crime prevention and recommendation systems, to address business problems that need data models to represent complex interactions. GDS allows insight to be derived from these structures beyond the expression of queries and visualizations to build predictive and prescriptive analytical models. GDS includes a wide variety of graph machine learning models that work directly on graph-structured data for node and link prediction.

Business Impact

Extending data science and machine learning techniques with GDS opens opportunities to tackle complex challenges where network effects and relationships cannot be easily modeled using tabular data, but are often better indicators of predicting an outcome. Such challenges are present in many industries. GDS used in conjunction with other techniques is part of the emerging discipline of decision intelligence and forms part of an organization’s advanced analytics ecosystem.

Drivers

- Graph databases and knowledge graphs are increasingly used as a unifying layer for data access and usage for downstream applications.
- Academia and industry research centers have made significant advancement in the creation and implementation of graph data science and machine learning algorithms.
- Cloud computing has reduced the burden on organizations to set up and maintain the infrastructure required to run complex graph workloads and has also made specialist hardware accessible that is suited to graph algorithms.
- GDS libraries, both open-source and commercial, enable access to GDS techniques for data scientists using familiar languages and tooling.
- Low-code applications enable graph algorithms and machine learning to be done by domain experts and citizen data scientists.
- Technology and digital native companies such as Google, Uber, Pinterest and Amazon are pioneering the use of graph machine learning and publicizing the value they bring.
- GDS applied to knowledge graphs can provide a level of explainability to predictive models that existing methods cannot provide.

Obstacles

- Expertise is required to understand the various algorithms in the GDS domain and when to apply the appropriate technique to a business problem.
- Awareness of graphs as a data representation solution is not widespread among business leaders and data scientists, who tend to stick to traditional approaches or off-the-shelf solutions.
- Some machine learning tasks on very large graphs still require a significant amount of compute infrastructure and require the manipulation and preprocessing of the graph into the correct structure.
- Operationalization platforms for models built using GDS algorithms are immature compared to traditional DSML platforms.

User Recommendations

- Identify business problems that have potential to be solved using graphs by engaging with domain experts and assessing the amount of data integration, processing and analytical workloads that can be optimized.
- Dedicate time for data scientists to explore graph frameworks and libraries, and create sandbox environments where ideas can be tested.
- Apply graph features to existing predictive models, where relationships are a key characteristic of the data and measure results.
- Create a team to educate and inform different audiences and stakeholders on the uses of graph machine learning and applicable use cases.

Sample Vendors

Amazon Web Services; ArangoDB; Graphistry; Kumo.ai; Neo4j; Oracle; TigerGraph; Virtualitics

Gartner Recommended Reading

3 Ways to Enhance AI With Graph Analytics and Machine Learning

How Graph Techniques Deliver Business Value

Reinforcement Learning

Analysis By: Peter Krensky, Shubhangi Vashisth

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Reinforcement learning (RL) is a type of machine learning (ML) where the learning system receives training only in terms of positive feedback (rewards) and negative feedback (punishments). During problem solving, the system fosters actions or situations so that the overall reward is maximized while minimizing punishments.

Why This Is Important

Some problems can best be solved with RL, especially when other ML approaches are not feasible due to a lack of labeled training data.

Business Impact

The primary potential of RL is in industrial control and design, marketing and advertising, recommendation systems, and gaming industries. The technology can lead to significant improvements in self-driving cars, robotics, vehicle routing, warehouse optimization, logistics, predictive maintenance and other industrial control scenarios.

Drivers

- Recent successes across various industries (For example, text summarization and machine translation, real-time bidding for marketing and advertising, creation of dynamic treatment regimes in healthcare, optimized design of chip layouts in manufacturing, and optimization of robotic players in gaming.)
- Commercial vendors launching new RL products and products with embedded RL
- Sustained data scientist interest in the RL framework because it involves much less training data and supervision than currently dominant supervised learning schemes
- Faster compute capabilities are enabling more application scenarios for RL
- Better simulation capability is also an enabler of RL scenarios
- Reinforcement Learning from Human Feedback (RLHF) in which feedback from an AI community or user group is used to train better models
- Increased attention, interest and potential recognition due to generative AI hype

Obstacles

- Limited RL capabilities offered by current data science and machine learning (DSML) platforms
- Often exceedingly high computational requirements
- Lack of good-enough simulations in many business situations
- Difficulty in designing the reward structure of the RL model for most business scenarios
- Often brittle or difficult-to-implement solutions with applicability in limited use cases
- Lack of staff with reinforcement learning experience
- Lack of explainability

User Recommendations

- Apply RL in use cases requiring frequent model retraining with traditional techniques, because RL can adapt to new environment and circumstances
- Apply RL when the business outcomes and constraints are clear but you lack sufficient labeled data to build robust ML models.
- Acquire special expertise or engage a service provider with risk management support. The application of RL is currently riskier than most traditional techniques.
- Leverage off-the-shelf capabilities available from major vendors in the market, and seek out embedded reinforcement learning.

Sample Vendors

AgileSoDA; Amazon Web Services (AWS); Dataiku; MathWorks; Microsoft; Pathmind; RISELab; TensorFlow

Gartner Recommended Reading

Innovation Insight for Generative AI

Innovation Insight: AI Simulation

Go Beyond Machine Learning and Leverage Other AI Approaches

Value Stream Management Platforms

Analysis By: Hassan Ennaciri, Akis Sklavounakis

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

A value stream management platform (VSMP) is a platform that seeks to optimize end-to-end product delivery and improve business outcomes. VSMPs are typically tool-agnostic. They connect to existing tools and ingest data from all phases of software product delivery — from customers' needs to value delivery. VSMPs help software engineering leaders identify and quantify opportunities to improve software product performance by optimizing cost, operating models, technology and processes.

Why This Is Important

As organizations scale their agile and DevOps practices, higher-level metrics that assess performance and efficiency of their product delivery are essential. VSMPs integrate with multiple data sources to provide DevOps-related telemetry. These insights enable stakeholders to make data-driven decisions in an agile manner and correct course as needed. The visualization capabilities of VSMPs help product teams analyze customer value metrics against the cost required to deliver that value.

Business Impact

VSMPs help organizations bridge the gap between business and IT by enabling stakeholders to align their priorities to focus on delivering customer value. VSMPs can provide CxOs with strategic views of product delivery health and pipelines, allowing them to make data-driven decisions about future product investments. These platforms also provide product teams with end-to-end visibility and insight into the flow of work to help them address constraints and improve delivery.

Drivers

- Improved software delivery with business priorities and objectives.
- Timely decision making driven by insights from data.
- Optimization of delivery flow through reduction of waste and elimination of bottlenecks.
- Visibility and mapping of end-to-end software delivery processes and identification of cross-team dependencies.
- Quality and velocity improvements of product deployments.
- More stringent governance, security and compliance requirements.

Obstacles

- VSMPs are not focused on continuous integration/continuous delivery (CI/CD) capabilities. Execution of the delivery pipeline requires use of a custom toolchain or DevOps platform.
- VSMPs require customization and data from tools used by multiple stakeholders in the organization, sometimes outside of software delivery. Collaboration with these key stakeholders to deliver the desired insights is paramount.
- VSMPs are still evolving and not all vendors have all the core capabilities.

User Recommendations

- Accelerate business outcomes by leveraging real-time, data-driven metrics and value stream insights provided by VSMPs.
- Leverage VSMPs' AI-powered analytics and insights to surface constraints, detect bottlenecks and improve flow.
- Build customized dashboards and views of product delivery for multiple stakeholders and leadership.
- Utilize VSMPs to assess the performance, quality and value of products, including development costs and ROI.
- Use VSMPs to gain a consolidated view of governance, security and compliance across all product lines.

Sample Vendors

Broadcom; ConnectALL; Digital.ai; HCLSoftware; IBM; OpenText; Opsera; Planview; Plutora; ServiceNow

Gartner Recommended Reading

Market Guide for Value Stream Management Platforms

Tools for Delivering Business Metrics to Software Engineering Teams

Market Guide for Value Stream Delivery Platforms

Use the Right Metrics in the Right Way for Enterprise Agile Delivery

Cloud Sustainability

Analysis By: Ed Anderson

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Cloud sustainability is the use of cloud services to achieve sustainability benefits within economic, environmental and social systems. As such, cloud sustainability refers to both the sustainable operation and delivery of cloud services by a cloud service provider, as well as the consumption and use of cloud services by organizations and individuals to achieve sustainability outcomes.

Why This Is Important

Cloud sustainability is a key digital technology supporting organizations in their use of technology to achieve their sustainability ambitions. Cloud computing models are well-suited to deliver sustainability benefits because of their ability to operate at scale using a shared services model, which results in efficient use of computing resources. Hyperscale cloud data centers can be physically located near renewable energy sources further extending their potential to lessen environmental impact.

Business Impact

Increasing attention and focus on environmental and social issues is motivating organizations to improve their sustainability posture. Pressure from customers, investors, partners, regulators, employees, and the public at large is motivating organizations to establish sustainability goals and to demonstrate sustainability outcomes. Cloud computing has great potential to improve sustainability outcomes through efficient operations and the delivery of cloud-based technology innovations.

Drivers

- Sustainability is a rising imperative for organizations across all industries and in all countries and regions around the world. Although sustainability encompasses environmental, social and economic factors, environmental sustainability receives the most attention.
- Corporate climate and decarbonization commitments are typically cascaded to individual business functions, including IT. Consequently, IT organizations are looking at all possible ways to implement such strategies, including cloud sustainability initiatives.
- Market data shows that customers, investors, regulators, citizens and employees increasingly value organizations with demonstrable commitments to sustainability.
- Sustainability investments correlate with operational efficiency. Most organizations operating in an increasingly sustainable fashion also recognize other benefits such as reduced spending on energy, reductions in waste and improvements in water use.
- Cloud providers, being among the world's largest data center operators, show strong commitments to cloud sustainability and are making demonstrable progress toward delivering sustainable cloud service offerings.
- Regulatory and legislative mandates for sustainability are increasingly common across regions and industries. The use of cloud services and other digital technologies will help organizations comply with future regulatory reporting requirements.

Obstacles

- Sustainability definitions, metrics and reporting standards are inconsistent, varying by region and industry. Defining, tracking and reporting sustainability performance is complex for most organizations.
- Cloud providers claim to have made great strides in offering sustainable cloud solutions, but these claims are often difficult to verify and contribute to potential "greenwashing." The lack of sustainability reporting standards makes it difficult to interpret and validate provider claims.
- Achieving cloud sustainability outcomes is a shared responsibility between the cloud provider and the customer. Cloud providers must demonstrate sustainable cloud operations, and cloud consumers must employ sustainability practices in their use of cloud services.
- Renewable energy is a key enabler of cloud sustainability and yet there is insufficient capacity to generate and store the energy required to meet the needs of the world's cloud service offerings.

User Recommendations

- Establish internal sustainability goals including specific metrics and sustainability outcomes by doing a materiality assessment to determine which sustainability outcomes are most important to your organization.
- Determine the role cloud sustainability will play in the achievement of sustainability outcomes. Build internal credibility for cloud sustainability by ensuring that the sustainability benefits of specific cloud service offerings are independently validated.
- Engage relevant executives and other internal stakeholders proactively that are tasked with creating and achieving sustainability goals. Establish credible metrics for measuring and reporting cloud sustainability outcomes.
- Look to cloud providers and other experts, including IT service providers, for best practices in operating and consulting cloud services in a sustainable manner.

Sample Vendors

Alibaba Cloud; Amazon Web Services; Google; IBM; Microsoft; Oracle; Salesforce; SAP; Scaleway; VMware

Gartner Recommended Reading

Executive Leadership: Sustainability Primer for 2023

Quick Answer: How Green Are Public Cloud Providers?

Build an Environmental Cloud Sustainability Strategy

Make Sure Technology Helps More Than Hurts Sustainability

Sustainability: A Customer Priority and Provider Imperative

Homomorphic Encryption

Analysis By: Mark Horvath, Bart Willemsen

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Homomorphic encryption (HE) uses algorithms to enable computations with encrypted data. Partial HE (PHE) supports only limited use cases, such as subtraction and addition, but with little performance impact. Fully homomorphic encryption (FHE) supports a wider range of repeatable and arbitrary mathematical operations; however, it worsens performance.

Why This Is Important

HE offers an unparalleled advance in privacy and confidential data processing, although this is largely at the database level. Benefits include the ability to:

- Perform analytics on data while in an encrypted state, so that the processor never sees the data in the clear, yet delivers accurate results.
- Share and pool data among competitors.
- Share all or part of users' data, while protecting their privacy.
- Systems based on lattice encryption, which are quantum-safe.

Business Impact

Even in restricted form (PHE), HE enables businesses to use data, send it to others for processing and return accurate results, without fear it will be lost, compromised or stolen. Any data intercepted by a malicious actor is encrypted and unreadable, even by the coming generation of quantum computers.

Applications include:

- Encrypted search
- Data analytics
- Machine learning (ML) model training
- Multiparty computing
- Securing, long-term record storage, without concerns about unauthorized decryption

Drivers

- The enhanced enforcement of data residency restrictions worldwide is forcing organizations to protect data in use, rather than only when it is in transit or at rest.
- Globally maturing privacy and data protection legislative frameworks demand that more-precise attention be paid to sensitive data. As a result, data pooling, sharing and cross-entity analysis use cases increasingly benefit from forward-looking and sustainable technologies, such as HE.
- Aside from primarily financial use cases (e.g., cross-entity fraud analytics), other industries can benefit as well. One example is the healthcare industry, where analysis of sensitive data across various entities happens often with data protected while in use.
- Solving issues of trust and cooperation with secure multiparty computation (sMPC) will benefit internal and external protection of data.
- The oncoming availability of quantum computing (QC), as highlighted by NIST and the Canadian Forum for Digital Infrastructure Resilience, threatens to compromise the confidentiality of almost all data. This includes digital communication previously considered protected by conventional cryptography. For example, there are signals that malicious actors may retain exfiltrated encrypted data in expectation of the ability to decrypt it years later and re-engage with victims for extortion and ransom demands. Timely adoption of HE in data protection will sustainably protect data, even when previously compromised in (conventionally) encrypted form.

Obstacles

- The application of various forms of HE to daily use cases leads to a degree of complexity, slows operations and requires highly specialized staff.
- The market's unfamiliarity with this technology stands in the way of speedy adoption.
- Although PHE can be a Turing-complete implementation, which means an arbitrary set of instructions could be executed, no vendor has a robust implementation that exploits this capability.
- Some scenarios will never be a good match for HE — for example, those that require security in components beyond analytics and processing, such as production databases and proprietary algorithms.

User Recommendations

- Brainstorm opportunities with your technical and executive teams. For example, come up with a list of five to 10 use cases for HE to improve the adoption of core solutions.
- Treat potential HE projects as experiments, keeping in mind the early stage of the technology's development and the significantly not-real-time nature of HE products. Consider these experiments proofs of concept (PoCs) to build experience, until the technology matures.
- Continue with existing security controls. HE does not necessarily negate the need for other security controls, observance of data residency requirements or access control.
- Assess the core benefits of using HE in combination with other quantum-safe or privacy-enhancing computation techniques.
- Integrate in-use protection via forms of HE into messaging and third-party analytics services.
- Assess the merits of piloting HE by using a vendor's solution, which could offer functionality without the time investment associated with a custom solution.

Sample Vendors

CryptoLab; Duality; Enveil; IBM; Inpher; IXUP; LiveRamp; Lorica; Ziroh Labs

Gartner Recommended Reading

Emerging Technologies: Homomorphic Encryption for Data Sharing With Privacy

Emerging Technologies and Trends Impact Radar: Security

Three Critical Use Cases for Privacy-Enhancing Computation Techniques

What Executives Need to Do to Support the Responsible Use of AI

Achieving Data Security Through Privacy-Enhanced Computation Techniques

Internal Developer Portal

Analysis By: Manjunath Bhat

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Internal developer portals serve as the interface that enables self-service discovery and access to resources in complex, cloud-native software development environments. They can include software catalogs, scorecards to benchmark software quality, scaffold templates, product documentation, plug-ins for extensibility and automation workflows. Developer portals help improve developer productivity, operational efficiency and enhance governance by providing shared visibility across multiple teams.

Why This Is Important

Internal developer portals help software developers navigate infrastructure complexity, understand service interdependencies and enable faster release cadence in at least three ways. First, they serve as a common viewpoint for multiple teams of developers. Second, they provide developers with self-service access to underlying platform components and environments. Third, they provide a centralized place to score applications and measure progress against reliability and security requirements.

Business Impact

Developer portals can have the following business impacts:

- Developer experience and productivity: Help development teams improve their delivery cadence by improving developer experience, reducing cognitive load and shortening the onboarding time.
- Reliability and resilience: Aim to provide visibility to application health and include scorecards to assess their production readiness.
- Security and governance: Include prebuilt toolkits, templates and curated libraries that help create “paved roads” with built-in compliance, security and audit policies.

Drivers

- Platform engineering: Organizations are adopting platform engineering principles and creating platform teams to scale cross-cutting capabilities across multiple development teams. Platform teams curate internal developer platforms to abstract away the complexity of siloed systems and processes. Internal developer portals serve as an interface through which developers can consume the capabilities of internal developer platforms.

- Backstage: Backstage is one of the first open-source frameworks for building developer portals. It was created at Spotify and is now a Cloud Native Computing Foundation (CNCF) incubating project. The thriving open-source community supporting Backstage has largely contributed to its enormous mind share and rapid adoption. Hundreds of organizations have adopted Backstage since it was open-sourced in 2020. Backstage's success continues to drive interest, momentum and competition in this space.

- Developer experience: With software at the core of all digital innovation today, a great developer experience that accelerates software development becomes a key competitive advantage. Therefore, software engineering leaders are increasingly focused on minimizing developer friction and frustration. The ability to curate and provide customizable, developer-friendly experiences within the developer portal and reign in complexity will drive their appeal for both product and platform teams.

- Innersource: To enable rapid innovation and facilitate greater collaboration and knowledge sharing, software engineering leaders are adopting innersource approaches to software development. However, innersource requires an easy way for other teams to discover and search for existing projects within their organization. This is why organizations adopting innersource are turning to internal developer portals to make projects available and discoverable by other teams. See InnerSource Portal.

Obstacles

- Prerequisites: The successful adoption of internal developer portals goes beyond deploying a tool and requires certain prerequisites to be in place. For example, application services and their dependencies must be organized with consistently defined metadata that helps track their usage, performance and team ownership.
- Absence of platform teams: A dedicated platform team to manage and evolve the portal as a product is necessary to ensure the portal meets desired objectives. The absence of a dedicated platform team, and more so, led by a platform product owner to manage the portal as a product results in a disconnect between developer expectations and the portal's capabilities.
- Lack of developer buy-in: Although the developer portal serves as the "window" to the underlying platform's capabilities, it should provide "paved roads" and not "forced marches" — portal use should remain the choice of the development team. Trying to force development workflows into organizationwide blueprints for building developer portals without involving developers is a recipe for failure.

User Recommendations

- Use internal developer portals to scale cross-cutting software engineering capabilities across multiple development teams and streamline the software delivery life cycle.
- Do not assume that internal developer portals are turnkey solutions — they require a lot of prerequisites to be in place and many cases involve several weeks of prework activities. For example, Backstage requires codification of service-related metadata in YAML files before the content shows up in the software catalog.
- Ensure that the platform team includes internal developer portals in their charter. Continuously innovate portal capabilities by appointing a platform product owner for the developer portal to manage its roadmap, gather feedback and market its capabilities.

Sample Vendors

Atlassian; Calibo; CodeNOW; configure8; Cortex; Mia-Platform; OpsLevel; Port; Roadie

Gartner Recommended Reading

Innovation Insight for Internal Developer Portals

Drive Innovation by Enabling Innersource

Adopt Platform Engineering to Improve the Developer Experience

Cool Vendors in Platform Engineering for Improving Developer Experience

Federated Machine Learning

Analysis By: Ben Yan, Svetlana Sicular, Pieter den Hamer, Mike Fang

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Federated machine learning aims at training a machine learning (ML) algorithm on multiple local datasets contained in local nodes without the explicit sharing of data samples. Federated ML helps to protect privacy, enables ML and specifically deep neural networks (DNNs) to use more data, resolves data transfer bottlenecks, and empowers collaborative learning for better accuracy.

Why This Is Important

Federated ML (FedML) highlights an important innovation in (re)training ML algorithms in a decentralized environment without disclosing sensitive business information. FedML enables more personalized experiences with local learning in smartphones, softbots, autonomous vehicles or IoT edge devices, and also facilitates organizations to build collaborative learning models across data silos.

Business Impact

FedML enables collaborative ML by sharing local model improvements at a central level, while keeping the data locally. It especially benefits the Internet of Things (IoT), cybersecurity, privacy, data monetization and data sharing in regulated industries. For example, the U.S. Department of Health and Human Services recently reported an average improvement of 16% and a 38% increase in generalization over local models, as a collaboration result of 20 institutes.

Drivers

- The proliferation of privacy regulations requiring protection of local data.
- With the increasing hype around edge AI, the data becomes distributed across multiple, heterogeneous edge devices and clouds. FedML allows organizations to keep the data in place.
- Data volumes are still growing rapidly, making it more challenging to collect and store big data centrally. This is especially pronounced in the IoT scenarios, where sensor data is collected on the devices and often there is no time or reason to pass it centrally.
- Due to scalability issues, excessive power consumption, connectivity and latency, we see a move toward edge infrastructure in the form of FedML.
- Organizations need collaboration with upstream and downstream partners to improve the overall operation efficiency.
- As large language model (LLM) evolves, research on federated LLM emerges so that a group of organizations could collaborate to train LLM together.
- Swarm (federated) learning is emerging as a promising approach in decentralized ML, uniting edge computing, peer-to-peer networking and coordination, enabled by blockchain.
- FedML is often combined with other privacy enhancing computation techniques as complete secured computing solutions.

Obstacles

- Building trust between organizations for collaborative learning models takes time.
- The incentive mechanism needs to be defined and agreed with all parties engaged to keep participants motivated and keep the FedML group in the long run.
- System and data heterogeneity requires a lot of coordination and standardization among systems to be fully functional.
- Enabling FedML requires a complete end-to-end infrastructure stack that integrates capabilities across DataOps, ModelOps, deployment and continuous tracking/retraining, necessitating a high degree of implementation maturity.
- Creating a new, more accurate and unbiased central model from local model improvements can be nontrivial, as the diversity or overlap between local learners and their data may be hard to assess and may vary greatly.
- FedML is still not widely known in the enterprise, as it lacks marketing on the vendor and researcher sides.
- Security and privacy validation concerns require additional steps.

User Recommendations

- Apply FedML to create and maintain decentralized smart services or products, while protecting the privacy of users and preventing the need to centrally collect massive amounts of data.
- Explore FedML use cases with upstream and downstream partners and look for opportunities to improve overall operation efficiency.
- Give a head start to decentral ML applications by deploying a common, centrally pretrained model while still providing personalization and contextualization by locally retraining the model based on local data and feedback.
- Enable continuous improvement of decentralized ML applications with collaborative learning by repeatedly collecting local model improvements to create a new, improved central model and then redeploying it for decentral usage and fine-tuning.
- Keep a central reference model to ensure “cognitive cohesion” across distributed models — that is, by avoiding decentralized models that veer off too far from its original purpose.

Sample Vendors

Alibaba Group; Devron; Ederlabs; F-Secure; Google; Intel; NVIDIA; Owkin; WeBank

Gartner Recommended Reading

Innovation Insight for Federated Machine Learning

Quick Answer: Why Is Federated Learning Prominent in China?

Explore Secured, Accurate and Green AI With Federated Machine Learning

Industry Cloud Platforms

Analysis By: Gregor Petri

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Industry cloud platforms address industry-relevant business outcomes by combining underlying SaaS, PaaS and IaaS services into a whole product offering with composable capabilities. These typically include an industry data fabric, a library of packaged business capabilities, composition tools and other platform innovations. IT leaders can use the composability of these platforms to gain the adaptability and agility their industries need to respond to accelerating disruption.

Why This Is Important

Cloud, software and service providers are launching industry cloud platforms (ICP) by combining SaaS, PaaS and IaaS offerings with industry-specific functionality and composable capabilities to create more compelling propositions for mainstream customers. Emerging industry cloud platforms are leveraging innovative approaches such as composable packaged business capabilities (PBCs), PBC marketplaces, data grids and fusion teams to accommodate faster change and platform adaptability.

Business Impact

Broader cloud adoption within enterprises requires more whole-product business solutions that enable defined industry scenarios and process models, rather than technology-oriented solutions that enterprises have to largely configure and integrate themselves. ICPs enable enterprises to adopt more holistic cloud strategies that span across established cloud service categories such as SaaS, PaaS and IaaS.

Drivers

- As the complexities of both business and technology continue to increase, enterprises are looking for more outcome-based engagements with their cloud providers. However, such outcomes must be flexible enough to be able to adapt to the changing circumstances.
- To be relevant and be able to resonate with enterprise audiences, such outcomes must be business relevant, specific, measurable and tangible — a goal that is easier achieved when approached in a specific industry context.
- Industry cloud platforms can create value for enterprises by bringing traditionally separately purchased solutions together in a composable and modular way. This simplifies the sourcing, implementation and integration process.
- Currently, industry cloud platforms are being initiated and created by various technology providers. In addition, we see some enterprises considering creating — often in collaboration with a technology provider — a dedicated industry cloud platform as the basis for a more autonomous industry ecosystem.
- Enterprises can gain business value from industry clouds through shared best practices; vertically specialized go-to-market and implementation teams; compliance of the infrastructure platform with industry-specific regulations.
- Value can also be gained through analytical capabilities to integrally mine the data from existing and new applications; industry-specific add-on functionality in front- and back-office enterprise applications; combined with collections of composable building blocks available from industry cloud marketplaces.
- Providers are on a pathway to creating whole-product offerings that cater directly to the established needs of vertical industry enterprises.

Obstacles

- Industry clouds are at risk of following the same path as classic government and community clouds where providers created difficult to support or slightly outdated copies of the original cloud with specific functionality.
- Industry cloud platforms can be overwhelming in terms of the wide breadth of functionality they potentially cover. Customers and providers must therefore be disciplined and not burn precious resources on fixing/replacing things that are not broken.
- Implementing an industry cloud platform must be approached as adding an exoskeleton, bringing new and improved capabilities rather than a vital organ transplant, replacing or repairing functionality that was already present.
- To reach their full potential, industry clouds will need to evolve into something best described as ecosystem clouds. Enterprises can leverage these ecosystems by participating in shared (business) processes, such as procurement, distribution, payment procession, and maybe even R&D and innovation.

User Recommendations

- Target ICPs to complement the existing application portfolio like an exoskeleton by introducing new capabilities that add significant value, rather than as full-scale replacements of largely already existing functionality with more up-to-date technology.
- Start building composability skills by engaging business technologists and fusion teams to create enterprisewide understanding and support for the ICP journey.
- Formulate rules for when to deploy ICP capabilities as a productive platform for optimization and modernization by improving existing processes, and when to actively recompose them for more differentiating transformation and innovation initiatives.

Sample Vendors

Amazon Web Services (AWS); Google; IBM; Infor; Microsoft; Oracle; Salesforce

Gartner Recommended Reading

Top Strategic Technology Trends for 2023: Industry Cloud Platforms

Presentation: Industry Cloud Platform Adoption by Vertical Industry

Analyzing Industry Cloud Offerings From CIPS Providers

Providers of Cloud Managed Services: Use Composable Industry Platforms to Productize Your Offerings

Changes and Emerging Needs Product Managers Must Address in the CIPS Market

At the Peak

AI TRiSM

Analysis By: Avivah Litan, Jeremy D'Hoinne, Bart Willemsen

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

AI trust, risk and security management (AI TRiSM) ensures AI model governance, trustworthiness, fairness, reliability, robustness, efficacy and data protection. AI TRiSM includes solutions and techniques for model interpretability and explainability, data and content anomaly detection, AI data protection, model operations and adversarial attack resistance.

Why This Is Important

AI models and applications deployed in production should be subject to protection mechanisms. Doing so ensures sustained value generation and acceptable use based on predetermined intentions. Accordingly, AI TRiSM is a framework that comprises a set of risk and security controls and trust enablers that helps enterprises govern and manage AI models and applications' life cycle — and accomplish business goals. The collateral benefit is enhanced compliance with forthcoming regulations, like the EU AI Act.

Business Impact

Organizations that do not consistently manage AI risks are exponentially inclined to experience adverse outcomes, such as project failures and breaches. Inaccurate, unethical or unintended AI outcomes, process errors and interference from malicious actors can result in security failures, financial and reputational loss or liability, and social harm. AI misperformance can also lead organizations to make suboptimal business decisions.

Drivers

- ChatGPT democratized third-party-provisioned generative AI and transformed how enterprises compete and do work. Accordingly, the risks associated with hosted, cloud-based generative AI applications are significant and rapidly evolving.

- Democratized, third-party-provisioned AI often poses considerable data confidentiality risks. This is because large, sensitive datasets used to train AI models are shared across organizations. Confidential data access must be carefully controlled to avoid adverse regulatory, commercial and reputational consequences.
- AI risk and security management imposes new operational requirements that are not fully understood and cannot be addressed by existing systems. New vendors are filling this gap.
- AI models and applications must be constantly monitored to ensure that implementations are compliant, fair and ethical. Risk management tools can identify and eliminate bias from training data and AI algorithms.
- AI model explainability must be constantly tested through model observations. Doing so ensures original explanations and interpretations of AI models remain active during model operations. If they don't, corrective actions must be taken.
- Detecting and stopping adversarial attacks on AI requires new methods that most enterprise security systems do not offer.
- Regulations for AI risk management — such as the EU AI Act and other regulatory frameworks in North America, China and India — are driving businesses to institute measures for managing AI model application risk. Such regulations define new compliance requirements organizations will have to meet on top of existing ones, like those pertaining to privacy protection.

Obstacles

- AI TRiSM is often an afterthought. Organizations generally don't consider it until models or applications are in production.
- Enterprises interfacing with hosted, large language models (LLMs) are missing native capabilities to automatically filter inputs and outputs — for example, confidential data policy violations or inaccurate information used for decision making. Also, enterprises must rely on vendor licensing agreements to ensure their confidential data remains private in the host environment.
- Once models and applications are in production, AI TRiSM becomes more challenging to retrofit to the AI workflow, thus creating inefficiencies and opening the process to potential risks.
- Most AI threats are not fully understood and not effectively addressed.

- AI TRiSM requires a cross-functional team, including legal, compliance, security, IT and data analytics staff, to establish common goals and use common frameworks — which is difficult to achieve.
- Although challenging, the integration of life cycle controls can be done with AI TRiSM.

User Recommendations

- Set up an organizational task force or dedicated unit to manage your AI TRiSM efforts. Include members who have a vested interest in your organization's AI projects.
- Work across your organization to effectively manage best-of-breed toolsets for enterprise-managed AI and applications that use hosted AI as part of a comprehensive AI TRiSM program.
- Avoid, to the extent possible, black-box models that stakeholders do not understand.
- Implement solutions that protect data used by AI models. Prepare to use different methods for different use cases and components.
- Establish data protection and privacy assurances in license agreements with vendors hosting LLM models — for example, Microsoft or OpenAI.
- Use enterprise-policy-driven content filtering for inputs and outputs to and from hosted models, such as LLMs.
- Incorporate risk management mechanisms into AI models and applications' design and operations. Constantly validate reliable and acceptable use cases.

Sample Vendors

AIShield; Arize AI; Arthur; Fiddler; ModelOp; Modzy; MOSTLY AI; Protopia AI; SolasAI; TrojAI

Gartner Recommended Reading

Use Gartner's MOST Framework for AI Trust and Risk Management

Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework

Cloud-Out to Edge

Analysis By: Ed Anderson, Bob Gill

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Cloud-out to edge describes an architectural construct where a centrally managed cloud environment, typically a hyperscale cloud, provides cloud service capabilities that are extended to edge environments. In a cloud-out to edge architecture, the cloud control plane, including security, identity and access management, governance, operations, programming models and interfaces, and other control elements, originate in the cloud and are then instantiated at the edge.

Why This Is Important

The move to public cloud drives centralization of operating processes, including the controls used to govern the environments. Cloud-out to edge is an architectural construct that supports the extension of public cloud control models to edge environments. Cloud-out to edge complements edge-in to cloud models. Cloud-out to edge is popular when organizations standardize IT operational control through centralized, public cloud environments.

Business Impact

IT environments are growing in complexity due to the expanding cloud use cases creating complexity, operational risks, and increased costs. Cloud-out to edge models extend cloud capabilities, including the cloud control plane, to other environments, including systems operating at the edge. Extending public cloud capabilities to edge environments can be a means to address the complexities of distributed, hybrid environments by unifying IT operations under a common operational framework.

Drivers

- Adoption of hyperscale public cloud services continues to increase. Gartner predicts continued growth in public cloud adoption with IT spending rates on public cloud services expected to grow almost 20% through 2027.
- Cloud operations have become critical for most organizations, driving increased investment in tools and skills in cloud management practices.
- Centralized, hyperscale cloud services are not well-suited for all application scenarios, particularly those better-suited to run at the edge. This creates an architectural divide between cloud and edge, which impacts operations, programming interfaces, security, identity and access, and application compatibility.
- Enterprise digitalization trends increasingly involve use cases and processes that operate in a distributed manner, driving the need for services deployed at edge locations.
- Distributed architectures, including edge computing and distributed cloud, can benefit from the distribution of cloud services from centralized environments to edge. Cloud-out to edge models can extend the cloud control plane to provide management, governance and oversight to edge environments.
- Standardizing technologies around cloud technologies and unifying operations using the cloud control plane can reduce complexity and help manage costs.

Obstacles

- Cloud-out to edge assumes a centralized cloud system, which may not exist in organizations that are still maturing their cloud strategy.
- Cloud-out to edge assumes the standardization of architecture, technologies and operational control using a centralized cloud service. This approach can increase dependence on a single cloud provider.
- Cloud-out to edge implementations may intersect with existing edge services, including operational technology, which may be managed outside the IT domain and funded by non-IT budgets.
- Multicloud strategies, which are common with most organizations, may conflict with the cloud-out to edge approach. Cloud-out to edge typically drives unification of technology, architecture and control to a single cloud environment, which may conflict with an organization's desired multicloud approach.
- Cloud provider offerings purported to support cloud-out to edge implementations are still maturing, and often don't deliver the full benefits of distributed cloud approaches.

User Recommendations

- Focus on the needs of use cases operating at the edge to determine whether an edge-in to cloud or a cloud-out to edge approach will work best.
- Establish a comprehensive cloud and edge strategy to guide cloud-out to edge and edge-in to cloud implementations. Let business value and operational benefits lead cloud and edge decisions.
- Build strong, centralized cloud operating capabilities before pursuing cloud-out to edge strategies. Cloud competencies will be critical to achieving success in cloud-out to edge implementations.
- Assess the risks and benefits of a cloud-out to edge approach, particularly if you have a stated multicloud strategy.
- Seek expert help from system integrators and managed service providers with expertise in both cloud and edge environments.

Sample Vendors

Alibaba Cloud; Amazon Web Services; Google; IBM; Microsoft; Oracle; VMware

Gartner Recommended Reading

I&O Platforms Primer for 2023

Market Guide for Edge Computing

Quick Answer: How to Make the Right Choice Between Hyperconverged, Traditional and Distributed Cloud Infrastructure

Distributed Cloud: Does the Hype Live Up to Reality?

Emerging Tech: Hyperscale Edge Enables Integrated Edge Infrastructure and Platform Services

Open-Source Program Office

Analysis By: Nitish Tyagi, Mark O'Neill, Mark Driver, Arun Chandrasekaran

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

An open-source program office (OSPO) is the center of competency to build strategies for governing, managing, promoting and efficiently using open-source software (OSS) and open-source data or models. The OSPO includes members from application delivery, legal, IT security, procurement and product management.

Why This Is Important

OSS is the backbone of digital innovation, with more than 95% organizations using it. However, ad hoc usage of OSS can lead to legal, security and viability risks. To manage these risks, OSPO or similar programs adoption has increased by 50% (see TODO's OSPO Survey 2022). Many large enterprises such as Alibaba, Apple, Bosch and Capital One have done so. OSPOs build cohesive strategies to efficiently use and promote OSS (see A CTO's Guide to Open-Source Software: Answering the Top 10 FAQs).

Business Impact

OSS drives accelerated software development, innovation, flexibility, cost savings and talent retention. A well-run OSPO ensures efficient consumption of OSS, implements effective governance policies, facilitates contributions and communicates the value of OSS to stakeholders. Embracing OSS, particularly contributing to or maintaining an OSS project, positively affects the retention of employees, especially developers. As the OSPO matures, it becomes a strategic partner in all technology decisions.

Drivers

- OSS is ubiquitous, but enterprises with an ad hoc approach to OSS have limited visibility into where and how OSS is used within their technology stack. An OSPO provides a strategic approach and required visibility by using correct tools.
- Poorly governed use of OSS, without proper assessment, exposes an enterprise to security, legal and viability risks. For example, ungoverned use of OSS components may violate licensing terms or infringe upon intellectual property rights, or OSS with a weak community may incur huge technical debt. An OSPO is responsible for governing the use of OSS.
- Software developers wish to contribute to open-source projects that they use in their day-to-day work. Actively supporting contributions to OSS projects helps talent acquisition and retention and ensures the longevity and relevance of the OSS your organization uses.
- Establishing the source and provenance of software components is essential, as legal requirements for software bills of materials grow (see Executive Order on Improving the Nation's Cybersecurity).

Obstacles

- The lack of funding and executive sponsorship is the biggest reason for the limited adoption of OSPOs. Establishing the correct metrics to predict the success of an OSPO is often challenging for organizations.
- Silos between different teams make it difficult for an OSPO to drive collaboration.
- Total cost of ownership is always attached with using an OSS project, but costs are often ignored when organizations do not plan their use of OSS.
- A lack of self-service tooling and automation in applying the policies related to consumption and publication can delay the software development life cycle.

User Recommendations

- Establish an enterprisewide OSPO and shift from ad hoc use of OSS to strategic use of OSS by building policies and processes to govern, assess and manage OSS. Fill up the core OSPO roles and include members from different stakeholders groups, such as application delivery, security, legal and procurement.
- Define the correct set of metrics such as developer productivity, and hiring and retention rates to measure the OSPO's success. Formulate and enforce a governance policy for consumption, contribution and creation of OSS by building an OSS governance committee.
- Work with platform engineering teams to provide the correct set of tools for artifact repository, security, issue tracking, continuous integration, continuous development, collaboration and knowledge management.
- Evaluate your OSPO's maturity level and execute strategies to gradually promote it to the next level as appropriate.

Sample Vendors

Bitergia; Cloud Native Computing Foundation (CNCF); Linux Foundation; TODO Group

Gartner Recommended Reading

Best Practices for Setting Up an Open-Source Program Office

A CTO's Guide to Open-Source Software: Answering the Top 10 FAQs

How to Create and Enforce a Governance Policy for Open-Source Software

How Software Engineering Leaders Can Mitigate Software Supply Chain Security Risks

Video Spotlight: Unlocking the Value of Open Source at Fannie Mae

WebAssembly (Wasm)

Analysis By: Oleksandr Matvitskyy, Gregg Siegfried

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

WebAssembly (Wasm) is a lightweight virtual stack machine and binary code format designed to support secure, high-performance applications on webpages. A growing number of programming languages can generate Wasm as a target, and applications beyond the web are becoming more common. Nonbrowser use cases range from Lua-like application extensibility mechanisms to server-side application services, as an alternative to containers or as a platform for serverless and edge applications.

Why This Is Important

Wasm has potential to disrupt runtime environments like VMs and containers by improving software portability, efficiency, performance and security. As a W3C standard, developed in partnership with vendors, web browsers support it today. The server-side Wasm ecosystem is emergent, with standardization via the CNCF and Bytecode Alliance. One of the co-founders of Docker has been quoted as saying, “If WASM+WASI existed in 2008, we wouldn’t have needed to create Docker. That’s how important it is.”

Business Impact

Similar to the benefits of Java VM, container environments or public cloud infrastructure, Wasm benefits are associated with technology, so business impacts are aligned with technology transformations like application replatforming and rearchitecting. Wasm represents technology disruption of the application runtimes with new risks for incumbent application longevity, security and compliance.

Drivers

- Browser performance: Modern, browser-based UIs can be complex and heavy with substantial application and presentation logic delegated to the client side. This requires a runtime that is faster than the interpreted JavaScript VM and able to deliver native or near-native compute performance.
- Portability of code and browser limitations: The JavaScript VM can be used on both browser and server side. However, restricting browser-side implementation to JavaScript creates obstacles for developers proficient in other languages or code sharing between browser and server. Wasm allows development in most popular languages, for both the browser and server side. It also supports multiple processor architectures including ARM, and is compatible with the Kubernetes ecosystem.
- Edge computing: The need to deliver and execute latency-sensitive code closer to the user is increasingly a requirement for many modern workloads. Wasm is a near-perfect vehicle for meeting this type of requirement due to its compact packaging and very low resource requirements.
- Security: The capability model supported by the Wasm runtimes allows an extremely granular model for managing the “sandbox” within which the code executes that minimizes the attack surface. Unlike Java, Wasm is designed to be secure by default.
- Scalability: The startup time for Wasm applications is near instantaneous (below one millisecond). In a server-side use case, rather than keeping idle request handlers waiting for traffic, the request handlers are created at the time that the requests are received.
- Language flexibility: Many programming languages can compile into Wasm. Rust is a particularly popular choice, but support is available today for JavaScript, Go, Python and C/C++, among others.

Obstacles

- Developer tooling: Wasm represents lower abstraction level compared to the modern popular runtimes, so developers need improved tooling, component libraries and frameworks to keep development productivity high.
- Architecture and skills: While Wasm is designed for interoperability, it's not just a matter of switching it on or off when developers compile their code directly for Wasm. Usage of Wasm in the existing applications requires significant changes in application architecture and design.
- Security risks: Wasm supports much more granular control and can be safer than JavaScript engine in browser implementation. However, current browsers' DOM reliance on JavaScript is blocking the opportunities for addressing browser security concerns with Wasm implementation.
- Toolchain maturity: The DevOps toolchain for building, testing, deploying and releasing Wasm has not yet reached a level of maturity sufficient for enterprise use outside of the experimental.

User Recommendations

- Pilot the use of Wasm for performance-sensitive client-side software rather than defaulting to JavaScript. This will acquaint product teams with the differences in developer experience and the language/toolchain requirements for incorporating it into your stack.
- Prepare for transition to Wasm implementations by selecting platforms and frameworks that already support Wasm or have it on their roadmap. This can be an easy win for the organizations choosing to minimize the impact on application architecture and implementation.
- Explore the server-side Wasm ecosystem by encouraging a small team to prototype with the platforms and tools available from Cosmonic and/or Fermyon.

Sample Vendors

Cloud Native Computing Foundation (WasmEdge); Cosmonic; Dylibso; Fermyon; Google (Flutter with CanvasKit renderer, Node.js); Microsoft (Blazor)

API-Centric SaaS

Analysis By: Yefim Natis, Anne Thomas, Mark O'Neill

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

API-centric SaaS is a cloud application service designed with programmatic request/reply, or event-based, interfaces (APIs) as the primary method of access (instead of the traditional, and now optional, user interfaces). The strategic intent for API-centric SaaS is to contribute a set of business software components, packaged for use by advanced or business technologists as building blocks for composing custom application processes and services for end users.

Why This Is Important

API-centric SaaS serves as a foundation of creative innovation by business and software companies. It exposes modular business software as reusable building blocks in custom application development. Organizations create composite application processes and experiences that are more varied than relying entirely on in-house resources, and better targeted than relying entirely on the SaaS provider. Greater creativity in application engineering translates to business empowerment for faster, safer and more efficient innovation.

Business Impact

Business organizations equipped to use API-centric SaaS create new application experiences for their employees and customers, through composition of the new and prebuilt business software components, some sourced from multiple applications. They gain access to more impactful innovation to be more adaptive to the changing needs of the users and to better respond to competitive opportunities. Procurement of application services gradually becomes better matched to the consumed value.

Drivers

- Modern application design relies on cross-application integration and composition, compelling application vendors to deliver their business functionality, optionally or primarily, equipped for programmatic access.
- The growing popularity of the “headless” SaaS architecture in digital commerce provides the acceleration of creative innovation when using the modular API-first application design, changing the users’ assessment criteria for SaaS to favor the support of composability.
- The technology and skills for integration, including management of APIs, are widespread, promoting increasingly advanced use of programmatic interfaces to business applications.
- The demands for advanced customization of application experience in business organizations have evolved to the point that SaaS vendors must allow rearrangement of their business functionality by their customers. API-centric SaaS serves that purpose.
- Many older applications are increasingly accessed via APIs to include them in the modernization and innovation of organizations’ IT. This prepares organizations’ skills and technologies to include API-centric SaaS capabilities into their software engineering practices.
- Business application design has become significantly partitioned into the back-end functionality with its APIs and the front-end multiexperience, each side using different tools and design expertise. Some business-oriented application vendors find it convenient to concentrate on the back-end data and business logic, and leave the finalized user experience to separate teams, including the customer’s own developers.

Obstacles

- API-centric SaaS is a relatively new phenomenon. Both SaaS vendors and business developers may lack the required skills and tools.
- The best practices for pricing and procurement of API-centric SaaS are not well-developed, delaying adoption or increasing its costs. The pricing of some occasional use of APIs is common (and expensive) and does not match the use practices of API-first application products.
- Using multisourced API-centric components for assembling new application processes and experiences requires some integration work that may not be supported in selected composition tools. This requires advanced software engineering skills and delays adoption of API-centric SaaS by mainstream organizations.
- Reduced or absent user interfaces packaged with an API-centric SaaS assume and require that the customer implement their own differentiated application and user experience. What is a welcome opportunity for innovation for some can be a burden to others, delaying adoption of API-centric SaaS.

User Recommendations

- Build the tools and skills of API management that recognize the added requirements to govern access to imported third-party APIs.
- Give preference to SaaS offerings that expose and price more of their business functionality as APIs and/or event streams.
- Plan for the increasing use of composition and integration of API-centric business software in the design and delivery of application services, processes and experiences.
- Ensure clean API-based separation of the back-end business logic and the front-end user experience in most enterprise applications, to maximize the long-term benefits of adopted API-centric SaaS.
- Give preference to application platform offerings that are well-equipped for managed access to external APIs and event sources.
- Practice use and governance of APIs and event streams in preparation for greater adoption of API-centric SaaS.
- Watch for opportunities to experiment with a new business model by offering some of your business functionality packaged as priced API products or services.

Sample Vendors

Algolia; Alloy; Clearbit; Cloudinary; Lob; MessageBird; Plaid; Strapi; Stripe; Twilio

Gartner Recommended Reading

Accelerate Digital Transformation With an API-Centric Architecture for Enterprise Applications

How to Successfully Implement API-First Integration

Partner With Product Managers to Ensure the Success of API-Based Products

Banking Product Leader Insight: Think Beyond APIs to Address Composability

Quick Answer: What GMs Need to Know About the Composable Future of Applications

AI-Augmented Software Engineering

Analysis By: Arun Batchu, Hema Nair, Oleksandr Matvitskyy

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

The use of artificial intelligence (AI) technologies (e.g., machine learning [ML] and natural language processing [NLP]) to help software engineers create, deliver and maintain applications is designated AI-augmented software engineering (AIASE). This is integrated with engineers' existing tools to provide real-time, intelligent feedback and suggestions.

Why This Is Important

Today's software development life cycle includes such routine and repetitive tasks as boilerplate functional and unit-test code and docstrings, which AIASE tools automate. AI-powered automation enables software engineers to focus their time, energy and creativity on such high-value activities as feature development. Emerging AI tools discover the configurations that meet operational goals. Software builders who use these tools remain productive and engaged, and they stay longer in their jobs.

Business Impact

AIASE accelerates application delivery and allocates software engineering capacity to business initiatives with high priority, complexity and uncertainty, helping quality teams develop self-healing tests and nonobvious code paths. These tools automatically generate test scenarios previously created manually by testers, and detect test scenarios often missed by test teams. AIASE tools detect issues with code security, consistency or maintainability and offer fixes.

Drivers

Demand drivers include:

- The increasing complexity of software systems to be engineered
- Increasing demand for developers to deliver high-quality code faster
- Increasing numbers of application development security attacks

- Optimizing operational costs

Technology solution drivers include:

- The application of AI models to prevent application vulnerabilities by detecting static code and runtime attack patterns
- The increasing impact of software development on business models
- The application of large language models to software code
- The application of deep-learning models to software operations

Obstacles

- Hype about the innovation has caused misunderstandings and unrealistic expectations about the benefits of AIASE.
- There is a lack of deep comprehension of generated artifacts.
- There is limited awareness about production-ready tools.
- Software engineers who fear job obsolescence have shown resistance.
- There is a lack of transparency and provenance of data used for model training.
- Uneven, fragmented solutions that automate only some of the tasks in the software development life cycle (SDLC).
- AI skills such as prompt engineering, training, tuning, maintaining and troubleshooting models.
- High model training and inference costs at scale.
- Intellectual property risks stemming from models trained on nonpermissive licensed code.
- Privacy concerns stemming from code, and associated proprietary data leaking as training data for AI models.
- Technical employees' fear of jobs being automated by AI.

User Recommendations

- Pilot, measure and roll out tools only if there are clear gains.
- Verify the maintainability of AI-generated artifacts, including executable requirements, code, tests and scripts.
- Track this rapidly evolving and highly impactful market to identify new products that minimize development toil and improve the experience of software engineers, such as those that ease security and site operations burden.
- Reassure software engineers that AIASE is an augmentation toolset for human engineers, not a replacement.
- Pick providers (including open-source vendors) that supply visibility to training data and transparency on how the model was trained.
- Establish the correct set of metrics, such as new release frequency and ROI, to measure the success of AIASE.

Sample Vendors

Akamas; Amazon Web Services; Diffblue; Google; IBM; Microsoft; OpenAI; SeaLights; Sedai; Snyk

Gartner Recommended Reading

Innovation Insight for ML-Powered Coding Assistants

Infographic: Artificial Intelligence Use-Case Prism for Software Development and Testing

Market Guide for AI-Augmented Software Testing Tools

Generative AI

Analysis By: Svetlana Sicular, Brian Burke

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Definition:

Generative AI technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. Generative AI has profound business impacts, including on content discovery, creation, authenticity and regulations; automation of human work; and customer and employee experiences.

Why This Is Important

Generative AI exploration is accelerating, thanks to the popularity of Stable Diffusion, Midjourney, ChatGPT and large language models. End-user organizations in most industries aggressively experiment with generative AI. Technology vendors form generative AI groups to prioritize delivery of generative-AI-enabled applications and tools. Numerous startups have emerged in 2023 to innovate with generative AI, and we expect this to grow. Some governments are evaluating the impacts of generative AI and preparing to introduce regulations.

Business Impact

Most technology products and services will incorporate generative AI capabilities in the next 12 months, introducing conversational ways of creating and communicating with technologies, leading to their democratization. Generative AI will progress rapidly in industry verticals, scientific discovery and technology commercialization. Sadly, it will also become a security and societal threat when used for nefarious purposes. Responsible AI, trust and security will be necessary for safe exploitation of generative AI.

Drivers

- The hype around generative AI is accelerating. Currently, ChatGPT is the most hyped technology. It relies on generative foundation models, also called “transformers.”
- New foundation models and their new versions, sizes and capabilities are rapidly coming to market. Transformers keep making an impact on language, images, molecular design and computer code generation. They can combine concepts, attributes and styles, creating original images, video and art from a text description or translating audio to different voices and languages.
- Generative adversarial networks, variational autoencoders, autoregressive models and zero-/one-/few-shot learning have been rapidly improving generative modeling while reducing the need for training data.
- Machine learning (ML) and natural language processing platforms are adding generative AI capabilities for reusability of generative models, making them accessible to AI teams.
- Industry applications of generative AI are growing. In healthcare, generative AI creates medical images that depict disease development. In consumer goods, it generates catalogs. In e-commerce, it helps customers “try on” makeup and outfits. In manufacturing, quality inspection uses synthetic data. In semiconductors, generative AI accelerates chip design. Life sciences companies apply generative AI to speed up drug development. Generative AI helps innovate product development through digital twins. It helps create new materials targeting specific properties to optimize catalysts, agrochemicals, fragrances and flavors.
- Generative AI reaches creative work in marketing, design, music, architecture and content. Content creation and improvement in text, images, video and sound enable personalized copywriting, noise cancellation and visual effects in videoconferencing.
- Synthetic data draws enterprises’ attention by helping to augment scarce data, mitigate bias or preserve data privacy. It boosts the accuracy of brain tumor surgery.
- Generative AI will disrupt software coding. Combined with development automation techniques, it can automate up to 30% of the programmers’ work.

Obstacles

- Democratization of generative AI uncovers new ethical and societal concerns. Government regulations may hinder generative AI research. Governments are currently soliciting input on AI safety measures.
- Hallucinations, factual errors, bias, a black-box nature and inexperience with a full AI life cycle preclude the use of generative AI for critical use cases.
- Reproducing generative AI results and finding references for information produced by general-purpose LLMs will be challenging in the near term.
- Low awareness of generative AI among security professionals causes incidents that could undermine generative AI adoption.
- Some vendors will use generative AI terminology to sell subpar “generative AI” solutions.
- Generative AI can be used for many nefarious purposes. Full and accurate detection of generated content, such as deepfakes, will remain challenging or impossible.
- The compute resources for training large, general-purpose foundation models are heavy and not affordable to most enterprises.
- Sustainability concerns about high energy consumption for training generative models are rising.

User Recommendations

- Identify initial use cases where you can improve your solutions with generative AI by relying on purchased capabilities or partnering with specialists. Consult vendor roadmaps to avoid developing similar solutions in-house.
- Pilot ML-powered coding assistants, with an eye toward fast rollouts, to maximize developer productivity.
- Use synthetic data to accelerate the development cycle and lessen regulatory concerns.
- Quantify the advantages and limitations of generative AI. Supply generative AI guidelines, as it requires skills, funds and caution. Weigh technical capabilities with ethical factors. Beware of subpar offerings that exploit the current hype.
- Mitigate generative AI risks by working with legal, security and fraud experts. Technical, institutional and political interventions will be necessary to fight AI's adversarial impacts. Start with data security guidelines.
- Optimize the cost and efficiency of AI solutions by employing composite AI approaches to combine generative AI with other AI techniques.

Sample Vendors

Adobe; Amazon; Anthropic; Google; Grammarly; Hugging Face; Huma.AI; Microsoft; OpenAI; Schrödinger

Gartner Recommended Reading

Innovation Insight for Generative AI

Emerging Tech Roundup: ChatGPT Hype Fuels Urgency for Advancing Conversational AI and Generative AI

Emerging Tech: Venture Capital Growth Insights for Generative AI

Emerging Tech: Generative AI Needs Focus on Accuracy and Veracity to Ensure Widespread B2B Adoption

ChatGPT Research Highlights

Cloud-Native

Analysis By: David Smith, Michael Warrilow

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Cloud-native refers to something created to optimally leverage or implement cloud characteristics. Those cloud characteristics are part of the original definition of cloud computing, and include capabilities delivered as a service. Cloud computing characteristics also include being scalable and elastic, shared, metered by use, service based, and ubiquitous by means of internet technologies.

Why This Is Important

Cloud-native is a popular term. Depending on its meaning, it can be described as taking full advantage of the cloud capabilities of a cloud provider, or using approaches pioneered in the cloud to deliver benefits wherever needed, via specific technologies such as containers. Cloud-native is not one thing, and there is a battle of ideas.

Business Impact

Cloud-native is a popular, hyped concept that aspires to attain and maximize the benefits of cloud computing; however, the realization of those benefits varies. For example, if a traditional, noncloud application is migrated to the cloud through a lift-and-shift approach, the application is unlikely to fully leverage cloud characteristics and deliver the maximum benefits. An application rewritten to take advantage of cloud capabilities is more likely to deliver the expected cloud outcomes.

Drivers

- The primary driver for cloud-native is the desire to “get the most out of the cloud.” The cloud itself means different things to different constituencies, so it’s not surprising that cloud-native means different things. What drives people to one or another of these approaches varies.
- Cloud-native can optimally leverage cloud technologies and benefits. The two most common meanings in use are contradictory. CSP-native is all about using native features and, therefore, locking yourself into a provider. Container-native focuses on containers, and may evolve into other technologies. This doesn’t guarantee portability, but is directionally consistent with the goal.
- There are multiple aspects to cloud-native, ranging from design to architectural to operational practices. Examples include LIFESPAR and the Twelve-Factor App (i.e., cloud-native application design) and DevOps (cloud-native operations).
- Cloud-native can be viewed on a continuum. It’s not a question of whether something is cloud-native or not; it’s the degree to which it is. The more it aligns with cloud characteristics, the more cloud-native it is.

Obstacles

- Cloud-native is confusing due to its many interpretations. It’s especially challenging with respect to hype, because confusion amplifies hype. The biggest obstacle is getting beyond the confusion to focus on desired outcomes.
- It is essential to be realistic about the portability that can be achieved and the cost. Otherwise, these features may not be used “with your eyes open,” and you may not be aware you are doing so.
- In cloud strategy efforts, principles are the most important component. Cloud-native and multicloud are often stated as principles in a cloud strategy. These principles can contradict each other, and require further explanation.
- Use of the term “cloud-native” requires clarification of which meaning is being used. This is a function of the hype surrounding cloud-native. Being clear about goals is key to optimally leveraging cloud-native. Assuming that containerizing an application will inherently make it cloud-native is an obstacle. We call this “container-native.”

User Recommendations

- Focus on the outcomes you want from using the cloud, rather than focusing purely on the definition of cloud-native. The more your use cases align with core cloud characteristics, the more likely you are to realize the benefits of using the cloud.
- Assess vendor claims about their cloud-native capabilities with skepticism. Vendors use the term “cloud-native” to promote their offerings, regardless of how cloud-native their offerings are.
- Ensure that the supporting tools, processes and operations support cloud characteristics when building or acquiring cloud-native applications or services. The value of cloud-native applications can be subverted when the approaches of the supporting elements are not cloud-native.
- Embrace services designed to bring you closer to cloud-native outcomes. These can include containers, microservices architecture, serverless design, functions and many platform-as-a-service (PaaS) services. However, using these technologies should be a means, not a goal.

Gartner Recommended Reading

The Cloud Strategy Cookbook, 2023

Infographic: Cloud-Native and Multicloud — Buzzwords or Key Principles in Your Cloud Strategy

A CTO’s Guide to Cloud-Native: Answering the Top 10 FAQs

Define and Understand New Cloud Terms to Succeed in the New Cloud Era

Appendixes

See the previous Hype Cycle: Hype Cycle for Emerging Technologies, 2022

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 2: Hype Cycle Phases

(Enlarged table in Appendix)

Phase	Definition
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Source: Gartner (August 2023)

Table 3: Benefit Ratings

Benefit Rating ↓	*Definition* ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (August 2023)

Table 4: Maturity Levels

(Enlarged table in Appendix)

Maturity Levels	Status	Products/Vendors
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (August 2023)

Evidence

We used as evidence for this research information from client inquiries, Gartner search analytics, Google trends and Gartner social media analysis of emerging technology topics in May 2023.

Document Revision History

Hype Cycle for Emerging Technologies, 2022 - 25 July 2022

Hype Cycle for Emerging Technologies, 2021 - 11 August 2021

Hype Cycle for Emerging Technologies, 2020 - 24 July 2020

Hype Cycle for Emerging Technologies, 2019 - 6 August 2019

Hype Cycle for Emerging Technologies, 2018 - 6 August 2018

Hype Cycle for Emerging Technologies, 2017 - 21 July 2017

Hype Cycle for Emerging Technologies, 2016 - 19 July 2016

Hype Cycle for Emerging Technologies, 2015 - 27 July 2015

Hype Cycle for Emerging Technologies, 2014 - 28 July 2014

Hype Cycle for Emerging Technologies, 2013 - 9 August 2013

Hype Cycle for Emerging Technologies, 2012 - 31 July 2012

Hype Cycle for Emerging Technologies, 2011 - 28 July 2011

Hype Cycle for Emerging Technologies, 2010 - 2 August 2010

Hype Cycle for Emerging Technologies, 2009 - 21 July 2009

Hype Cycle for Emerging Technologies, 2008 - 9 July 2008

Hype Cycle for Emerging Technologies, 2007 - 13 July 2007

Recommended by the Authors

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

Understanding Gartner's Hype Cycles

Tool: Create Your Own Hype Cycle With Gartner's Hype Cycle Builder

Assessing Emerging Technology Adoption Readiness

Table 1: Priority Matrix for Emerging Technologies, 2023

Benefit ↓	*Years to Mainstream Adoption*			
	Less Than 2 Years ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational		AI-Augmented Software Engineering Generative AI	Augmented FinOps Generative Cybersecurity AI Homomorphic Encryption Industry Cloud Platforms WebAssembly (Wasm)	Cybersecurity Mesh Architecture
High		AI TRiSM API-Centric SaaS Causal AI Cloud-Out to Edge Cloud Sustainability GitOps Internal Developer Portal Open-Source Program Office Postquantum Cryptography Value Stream Management Platforms	AI Simulation Cloud Development Environments Cloud-Native Federated Machine Learning Graph Data Science Reinforcement Learning	Neuro-Symbolic AI
Moderate				
Low				

Source: Gartner (August 2023)

Table 2: Hype Cycle Phases

Phase	*Definition*
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant media and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the innovation is pushed to its limits. The only enterprises making money are conference organizers and content publishers.
Trough of Disillusionment	Because the innovation does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the innovation's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the innovation are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the innovation to reach the Plateau of Productivity.

Phase ↓	Definition ↓

Source: Gartner (August 2023)

Table 3: Benefit Ratings

Benefit Rating ↓	Definition ↓
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (August 2023)

Table 4: Maturity Levels

Maturity Levels ↓	*Status* ↓	*Products/Vendors* ↓
Embryonic	In labs	None
Emerging	Commercialization by vendors Pilots and deployments by industry leaders	First generation High price Much customization
Adolescent	Maturing technology capabilities and process understanding Uptake beyond early adopters	Second generation Less customization
Early mainstream	Proven technology Vendors, technology and adoption rapidly evolving	Third generation More out-of-box methodologies
Mature mainstream	Robust technology Not much evolution in vendors or technology	Several dominant vendors
Legacy	Not appropriate for new developments Cost of migration constrains replacement	Maintenance revenue focus
Obsolete	Rarely used	Used/resale market only

Source: Gartner (August 2023)