Hype Cycle for Emerging Technologies, 2024

8 August 2024 - ID G00812275 - 107 min read

By Analyst(s): Arun Chandrasekaran, Jason Wong, Ankita Khilare

Initiatives:Technology Innovation and Strategy

Disruptive technologies hold great potential, as well as risks, for businesses and society. Technology innovation leaders should explore autonomous AI, boost developer productivity, empower with total experience, and deliver human-centric security and privacy programs.

Analysis

What You Need to Know

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

- **Autonomous AI:** AI's fast evolution is producing systems that can operate with minimal oversight, improve themselves and become effective at decision making in complex environments.
- **Developer productivity:** Technology is increasing developer productivity, enabling developers to deliver higher-quality products more quickly.
- **Total experience:** Intertwining customer experience, employee experience, multiexperience and user experience helps retain customers and employees, cultivating greater lifetime value from the relationships.
- **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 explore Gartner's recommendations for how to use them to drive competitive differentiation and efficiency.
- Examine technologies with transformational potential for your business capabilities, assessing 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 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 stage. Much uncertainty exists about how they will evolve. Emerging technologies present greater risks for deployment, but potentially significant benefits for early adopters.

Limited space means we've removed many technologies highlighted in the 2023 version. Those remain important and still feature in other Hype Cycles (see the Off the Hype Cycle section).

Themes in Emerging Technologies

In 2024, our emerging technology coverage focuses on autonomous AI, developer productivity, total experience, and human-centric security and privacy.

Autonomous AI: Autonomous AI is among the most advanced types of AI. Autonomous AI systems can operate with minimal human oversight. They seek to "understand" their environment, draw conclusions from it and adjust their actions accordingly. They can make decisions, purchase things and perform tasks, achieving goals in a range of environments as effectively as humans can. Systems that can perform any task a human can perform are beginning to move slowly from science fiction to reality.

Evaluate:

- AI supercomputing
- Artificial general intelligence
- Autonomous agents
- Generative AI
- Humanoid working robots
- Large action models
- Machine customers

- Multiagent systems
- Reinforcement learning

Developer productivity: Developer productivity is about more than writing code quickly. It's influenced by developers' effective communication and collaboration, and their feeling of energized focus, full involvement and enjoyment (being in the "flow state"). It also hinges on developers creating solutions that achieve valuable outcomes. Technology is revolutionizing the way developers design and deliver software, making them more productive than ever. It's also ensuring they deliver higher-quality products as quickly as possible. In addition, it's maximizing gains by improving developer satisfaction, collaboration and flow.

Examine:

- AI-augmented software engineering
- Cloud-native
- GitOps
- Internal developer portals
- Prompt engineering
- WebAssembly

Total experience: The benefits of connecting and empowering customers and employees, and removing friction for them, are obvious. Total experience is a strategy that creates superior shared experiences by intertwining customer experience, employee experience, multiexperience and user experience practices. It applies technology to address critical interactions to enable and embolden both customers and employees in their journeys. Its goal is to drive greater customer and employee confidence, satisfaction, loyalty and advocacy using digital and nondigital techniques.

Assess:

- 6G
- Digital twin of a customer
- Spatial computing

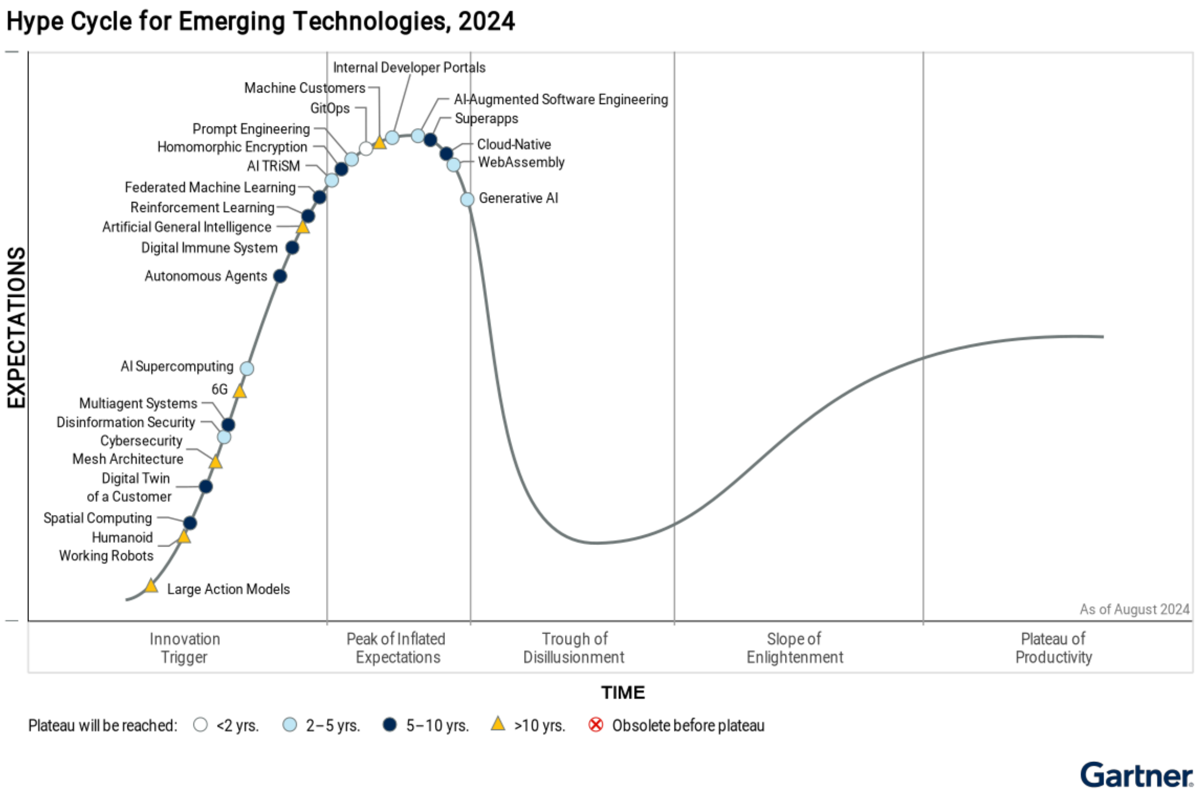
- Superapps

Human-centric security and privacy: Security practices too often rely on the premise that humans can behave in a completely safe and secure way. But when employees have to make a choice between security and business delivery, they often choose business delivery, sometimes bypassing too-stringent security controls. Organizations can become resilient by accepting humans' fallibility and implementing a human-centric security and privacy program. Such a program weaves a tight security and privacy fabric into the organization's digital design. New technologies enable organizations to create a culture of mutual trust and awareness of shared risks in decision making between many teams.

Explore:

- AI TRiSM
- Cybersecurity mesh architecture
- Digital immune system
- Disinformation security
- Federated machine learning
- Homomorphic encryption

Figure 1: Hype Cycle for Emerging Technologies, 2024



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.

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.

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.

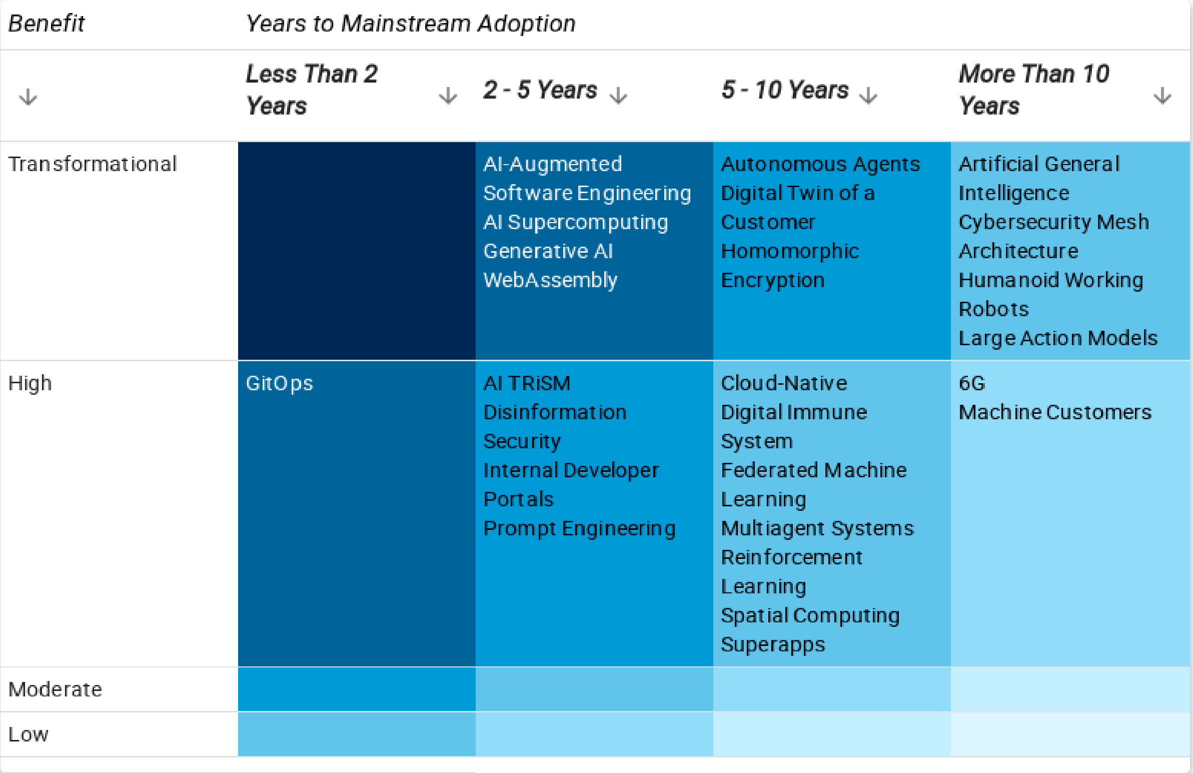
Using the Priority Matrix, technology innovation leaders should:

- Examine technologies that offer more significant near-term benefits that are both strategic and tactical.
- Explore technologies with longer-term benefits if they too are likely to offer strategic value.

Additionally, 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: Gartner's Hype Cycle Builder).

Table 1: Priority Matrix for Emerging Technologies, 2024

(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 AI Supercomputing Generative AI WebAssembly	Autonomous Agents Digital Twin of a Customer Homomorphic Encryption	Artificial General Intelligence Cybersecurity Mesh Architecture Humanoid Working Robots Large Action Models
High	GitOps	AI TRiSM Disinformation Security Internal Developer Portals Prompt Engineering	Cloud-Native Digital Immune System Federated Machine Learning Multiagent Systems Reinforcement Learning Spatial Computing Superapps	6G Machine Customers
Moderate				
Low				

Source: Gartner (August 2024)

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 2024 appear in the Hype Cycles indicated (information is correct at the time of writing):

- AI simulation — Hype Cycle for Artificial Intelligence, 2024; Hype Cycle for Generative AI, 2024; Hype Cycle for AI in Software Engineering
- API-centric SaaS — Renamed "headless SaaS" and appears in Hype Cycle for Cloud Computing, 2024; Hype Cycle for Cloud Platform Services, 2024; Hype Cycle for APIs, 2024; Hype Cycle for Banking Product and Service Innovation, 2024; Hype Cycle for Industry Cloud Platforms, 2024
- Augmented FinOps — Hype Cycle for Data Management, 2024; Hype Cycle for Operation Models, 2024; Hype Cycle for Site Reliability Engineering, 2024; Hype Cycle for Monitoring and Observability, 2024; Hype Cycle for Container Technology, 2024; Hype Cycle for Data, Analytics and AI Programs and Practices, 2024; Hype Cycle for IT Management Intelligence, 2024
- Causal AI — Hype Cycle for Artificial Intelligence, 2024; Hype Cycle for Finance AI and Advanced Analytics, 2024
- Cloud development environments — Hype Cycle for Cloud Platform Services, 2024; Hype Cycle for Software Engineering, 2024; Hype Cycle for Platform Engineering, 2024
- Cloud-out to edge — Hype Cycle for Oil and Gas, 2024
- Cloud sustainability — Hype Cycle for Infrastructure and Operations, 2024; Hype Cycle for Environmental Sustainability, 2024; Hype Cycle for Cloud Computing, 2024; Hype Cycle for Infrastructure Strategy, 2024; Hype Cycle for Operation Models, 2024

- Generative cybersecurity AI (renamed “cybersecurity AI assistants”) — Hype Cycle for Security Operations, 2024; Hype Cycle for K-12 Education, 2024; Hype Cycle for Cyber-Risk Management, 2024; Hype Cycle for Workload and Network Security, 2024; Hype Cycle for IT Management Intelligence, 2024; Hype Cycle for Emerging Technologies in the Communications Industry, 2024

- Graph data science — Hype Cycle for Data Science and Machine Learning, 2024; Hype Cycle for Banking Data, Analytics and AI, 2024

- Industry cloud platforms — Hype Cycle for Application Architecture and Integration, 2024; Hype Cycle for Cloud Computing, 2024; Hype Cycle for Cloud Platform Services, 2024; Hype Cycle for Higher Education, 2024; Hype Cycle for K-12 Education, 2024; Hype Cycle for Industry Cloud Platforms, 2024

- Neuro-symbolic AI — Hype Cycle for Artificial Intelligence, 2024; Hype Cycle for Finance AI and Advanced Analytics, 2024; Hype Cycle for AI in Software Engineering

- Open-source program office — Hype Cycle for Software Engineering, 2024; Hype Cycle for Platform Engineering, 2024; Hype Cycle for Open-Source Software, 2024

- Postquantum cryptography — Hype Cycle for Deep Technologies, 2024; Hype Cycle for Local Government, 2024; Hype Cycle for Defense, 2024; Hype Cycle for Data Security, 2024; Hype Cycle for Digital Banking Transformation, 2024

- Value stream management platforms — Hype Cycle for Infrastructure Platforms, 2024; Hype Cycle for Agile and DevOps, 2024

On the Rise

Large Action Models

Analysis By: Frank O'Connor

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Large action models (LAMs) are foundation models trained and optimized to identify and generate an action or set of actions that can be used to impact a target environment to meet a goal. For example, a large language model (LLM) solution can recommend a good restaurant to you; a LAM can book it for you.

Why This Is Important

By integrating LLMs with LAMs, users can express their goals in natural language. LAMs devise the required actions and deploy AI agents — software entities capable of autonomously executing tasks — to accomplish those goals across diverse digital or physical environments. LAMs have the potential to significantly disrupt the field of UI/user experience by providing interactions that are not just intuitive, flexible and efficient, but also highly personalized.

Business Impact

Industries likely to benefit from LAMs are automotive, education, healthcare, professional services, technology and software, with impacts including:

- Enhanced customer service experiences with more natural interactions, improving user satisfaction.
- Guided intelligent decision support solutions across different business functions.
- Delivery of personalized content and experiences as well as curated learning journeys.
- Better patient monitoring and diagnostic support, streamlining healthcare workflows.

Drivers

- **AI breakthroughs:** Recent breakthroughs in sophistication and utility of AI, such as generative pretrained transformers and LLMs, are enabling LAMs to better understand the intention behind human prompts and plan how to deliver outcomes that satisfy human prompts.

- **Multimodal understanding:** LAMs' ability to use diverse modalities like vision, audio and language enables more general and flexible AI assistants and agents. This allows automatic adaptation to changes in the workflow, user interface or API. This understanding enables LAMs to take high-level instructions and create advanced workflows without explicit programming, significantly reducing the development time and effort for automation.

- **Open-ended task capabilities:** Current AI models struggle with open-ended tasks requiring action sequences. LAMs could allow more complex and complete interactions by spanning agents to deliver more composable interactions that can achieve more wide-ranging goals.

- **Reinforcement learning integration:** LAMs provide a promising avenue to combine the pattern recognition abilities of LLMs with reinforcement learning for decision making and planning complex behaviors.

- **Autonomous systems:** Self-driving cars, drones, space exploration and other autonomous systems may employ LAMs to improve their ability to robustly execute complex objectives.

- **Robotics and physical world interaction:** LAMs could enable more capable and flexible robotic systems that can learn to execute complex action sequences from data, rather than relying solely on manually programmed routines. This has applications in markets such as manufacturing, logistics, healthcare and household assistance.

- **Virtual agents and environments:** Before being deployed in the real world, LAMs can first be developed and tested extensively in virtual environments and simulations. This allows safer iteration and scalable data collection for training the models.

Obstacles

- **Lack of trust:** Organizations are unsure if the human customer can trust the technology to accurately predict and execute tasks, and if the machine customer can trust the organization offering the service.

- **Data scarcity:** Training LAMs requires vast datasets of mapped action sequences and their corresponding goals across diverse environments.

- **Computational demands:** LAMs are computationally intensive to train due to the complexity of modeling action sequences, requiring immense computational resources.

- **Interpretability and oversight:** Action policies learned by LAMs may be opaque and have poor explainability, requiring mechanisms for human interpretability, oversight and control.

- **Accessibility:** While LAMs and AI agents have the potential to transform access for disabled people, care must be taken to ensure their interfaces are accessible and inclusive.

User Recommendations

- Start with focused, constrained applications and environments rather than aiming for general capabilities initially.

- Invest heavily in high-fidelity simulation infrastructure to enable scalable training and safe iterative development before real-world deployment.

- Build interpretability, human oversight and control mechanisms from the ground up to ensure LAM actions remain transparent and aligned with intended goals.

Sample Vendors

Covariant; MultiOn; rabbit; Toloka AI

Gartner Recommended Reading

Predicts 2024: The Future of Generative AI Technologies

Research Roundup for Generative AI

Innovation Insight: AI Agents

Humanoid Working Robots

Analysis By: Dwight Klappich

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Humanoid robots are designed and built to look like and mimic the human body in shape, function and locomotion. Humanoid robots imitate the human body and are powered by an AI-enabled brain. They have a head with sensors and cameras for sensing their environment; a body that houses the power and mechanicals; arms and hands/grippers for grasping, manipulating and carrying items; and legs for dynamic locomotion.

Why This Is Important

While there are many examples of highly productive mobile robots that use wheels, there have been few successful examples of robots that offered mobile manipulation (i.e., moving around and picking up and carrying things) and that could handle varied or unpredictable tasks and terrain. The next generation of humanoid working robots will combine sensory awareness with mobile manipulation and dynamic locomotion to perform productive work that was previously relegated to biological humans.

Business Impact

While today's mobile robots are proven and deliver the value expected, they lack the agility, versatility, and adaptability of humans. These robots typically do one or maybe two things well, while humans are nearly infinitely adaptable. This generation of humanoid robots is approaching the adaptability of the human workforce, where robots can flexibly support the needs of the business by dynamically moving from process to process and taking on new activities without special programming.

Drivers

- Until recently, there were no successful examples of humanoid robots that used legs and arms to navigate, pick up, move and manipulate items. Next-generation humanoid working robots will approach the level of adaptability that humans have, which will allow robots to be used and repurposed for many different activities without programming. This will allow the fluidity of work assignment expected of a company's human workforce.

- Next-generation humanoid robots, which are already hitting the market, will compete with the cost of labor for similar functions and will provide higher availability and reliability than the human workforce.

- Although robots lack fine motor skills and high-speed navigation, there are several activities that would benefit from the agility these robots offer. For example, the robot could do one task in the morning like loading totes or boxes on a conveyor and then could transition to pulling totes or boxes off the conveyor in the afternoon.

- Currently, activities like moving around and picking up individual items of varying sizes, shapes and orientations are difficult for robots but easy for humans. Consequently, humanoid robots will evolve during the next several years to address the limitations of previous generations of automation and will provide the necessary cost, flexibility, adaptability, scalability, utility and intelligence.

Obstacles

- While the technology is progressing rapidly, and there are numerous vendors competing for this emerging category of automation, this remains a nascent market, and we are years away from general-purpose humanoid robots that can approach the flexibility of humans.

- Given the mobile and varying types of work required in warehouses, the immobile, inflexible and programmed-to-do-one-thing architecture of traditional industrial robots is too limiting. Currently, robots are principally useful for addressing repeatable tasks with minimal variance from activity to activity or for moving things around from one location to another.

- As of now, industrial robots are slow compared to people or mobile robots. They lack the dexterity and fine motor control of humans, so the types of tasks they can perform are limited.

User Recommendations

- Explore how humanoid robots can be integrated into your processes. Supply chain operations with high-volume, predictable and consistent processes should consider the potential for humanoid robots and should not delay investing in them. Some of the current generations of intralogistics smart robots can help enhance or supplement the human workforce.
- Map your functional processes before pursuing humanoid robot companies to determine which, if any, of these processes can be addressed with current generations of robots, and which will benefit most from human-centered design.
- Identify processes that need high degrees of flexibility and adaptability, and focus on learning how your human workforce performs these tasks to understand what capabilities a humanoid robot will need to support.
- Develop a structured methodology for conducting effective proofs of concept while leading-edge companies are building robotics competency and collaboration centers (RCCC) to take on these roles.

Sample Vendors

Agility Robotics; Apptronik; Boston Dynamics; Figure; Fourier Intelligence; Tesla

Gartner Recommended Reading

Cool Vendors in Logistics Technology

Predicts 2024: Supply Chain Technology

Top Trends in Supply Chain Technology for 2024

Spatial Computing

Analysis By: Will Grant

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

Spatial computing is an emerging computing paradigm that combines physical and digital objects in a shared frame of reference, beyond screen-based displays. This involves spatial mapping and identification of people, places and things within the physical world as a foundation for anchoring digital content that intersects with the physical world's spatially anchored, indexed and organized content.

Why This Is Important

Spatial computing allows users to interact with digital content in a 3D space, making it a more immersive and intuitive experience. Given the potential of spatial computing to revolutionize sectors ranging from gaming and entertainment to healthcare and education, it is important for businesses to understand and adapt to this emerging technology.

Business Impact

Spatial computing has the potential to enhance productivity, collaboration and customer engagement across various applications. In general purpose end-user computing, it has the potential to increase productivity with "infinite" virtual displays and to make virtual meetings more engaging and interactive. Today, expertise is limited, as are protocols and tools to develop and experience spatial computing. However, enterprises are already developing and testing use cases across different areas.

Drivers

- **Technology:** Advancements in augmented reality (AR), virtual reality (VR), and AI technologies, such as head mounted displays (HMDs), spatial mapping, sensor fusion and AR cloud, have made it possible to create more realistic and immersive digital environments.

- **Consumer demand:** The increasing consumer demand for immersive and interactive experiences in sectors such as gaming, entertainment, and e-commerce is driving the adoption of spatial computing.

- **Specialized industry uptake:** In industries like healthcare and manufacturing, the need for sophisticated visualization tools for better decision making and increased efficiency is pushing the uptake of this technology.

- **Cellular data:** The proliferation of 5G and 6G technology, which provides the necessary speed and bandwidth for real-time interactions in spatial computing, is also a critical driver.

Obstacles

- **Standardization:** Providers establishing their own standards and protocols result in potentially incompatible methods for delivering a spatial computing environment and may potentially slow adoption.
- **Social isolation:** HMDs tend to isolate a person's perception from those around them (inhibiting eye contact, microexpressions, and so on). This could limit the impact of spatial computing in contexts that require deep human interaction.
- **Cost:** The high costs associated with AR and VR HMDs and digitizing assets are a significant barrier for many businesses and consumers scaling this technology.
- **Security:** Data privacy and security are causes for concern, given the immersive and interconnected nature of spatial computing and the vast amounts of personal data that can potentially be stored.
- **Compute power:** Technical issues, such as network latency and the need for high computational power, pose additional challenges.

User Recommendations

- Identify specific use cases where this technology can add value, which could be enhancing customer experiences, improving operational efficiency or facilitating remote collaboration.
- Prioritize spatial computing use cases in organizations that expand the utility and reach of their products and services.
- Invest in the necessary infrastructure, such as high-speed internet and compatible devices.
- Collaborate with experienced vendors to ensure a successful implementation of spatial computing. These vendors can provide the necessary expertise and support to overcome challenges and maximize benefits.
- Prioritize data privacy and security concerns associated with spatial computing and ensure that robust measures are in place.
- Stay abreast of spatial computing technologies and trends. This will enable organizations to leverage the full potential of this technology and stay ahead of the competition.

Sample Vendors

Apple; HTC; Magic Leap; Meta; Microsoft; Resight; Sony; Unity Technologies; VERSES

Gartner Recommended Reading

Quick Answer: What Is Spatial Computing?

Emerging Technologies: Tech Innovators in Augmented Reality — Spatial Web

Emerging Tech Impact Radar: The Metaverse

Emerging Tech: Head-Mounted Display Challenges Remain a Key Bottleneck for Metaverse Adoption

Emerging Tech: Accelerate Augmented Reality HMD Adoption in Enterprise Markets

Digital Twin of a Customer

Analysis By: Melissa Hilbert, Michelle DeClue

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Embryonic

Definition:

A digital twin of a customer (DToC) is a dynamic virtual mirror representation of a customer that organizations can use to simulate, as well as to emulate and anticipate, behavior and/or fit. Customers can be individuals, enterprise customer, personas, groups of people or machines.

Why This Is Important

DToCs help organizations better understand their customers and anticipate behavior given certain combinations of data and parameters (e.g., product, service). They increase efficiency and provide a personalized, empathetic service to customers, many of whose buying habits repeatedly change. Organizations can use a DToC to modify and enhance CX and support new digitalization efforts, products, services and opportunities. DToC can be an engine of transformation and disruption.

Business Impact

Today, organizations use digital twins to monitor how a product performs and to determine the next best action. Organizations can now use DToCs to simulate how a customer will react, given a specific set of ecosystem parameters, conditions, and control or input signals. DToCs can transform the way organizations selling products or services provide customers with better experiences, which results in increased revenue and lasting customer relationships.

Drivers

DToCs will help organizations drive revenue by:

- Gaining critical insights into customers.
- Increasing revenue by enabling new ways to serve or capture customers, as well as by facilitating new data-driven business models.
- Predicting and simulating behaviors with a view to making products, services, promotions and business campaigns more successful and reducing unnecessary costs of failure.
- Improving customer engagement, customer retention, customer lifetime value and company growth.
- Reducing churn, fraud, product failure and engagement abandonment.

DToCs will help customers:

- Reduce friction in interactions with the supplier organization throughout their journey.
- Increase positive outcomes, creating more personalization and value.
- Engage in curated experiences and concierge-like experiences specifically tailored to drive value for them.
- Protect privacy, with the ability to change what personal data is collected and how organizations use it.

Obstacles

- Privacy and cyber risk concerns for simulations not originally agreed to by the customer; increased compliance risk.
- Organizations need competency in machine learning algorithms and staff with data science skills to build or manage DToCs.
- Internal biases and concerns about a DToC's ability to drive revenue or reduce costs. Organizations will need a strategy based on use cases of how to create value.
- The technology behind digital twins has focused on organizations and products. A customer focus is emerging, and a lack of clear KPIs and other success measures limits the DToCs' potential use.
- Organizations need to establish trust with customers for them to agree to share their information. Customers will need transparency about what data is collected, how it will be used, and the privacy and data controls that will be applied. For B2B, they need to know the benefits, such as providing a more personalized experience, more relevant products or services, convenience, and exclusive offers.

User Recommendations

- Define clearly KPIs and specific objectives that can be measured to validate improved business outcomes, such as CX, demand forecastability or agility of responsiveness.
- Align activities with customers' privacy and cybersecurity concerns based on the availability of customer assets and establish a trust center to house these documents and expectations.
- Identify use cases for which DToCs could help deliver a better CX, and for which suitable data is available, by examining customer journeys and failure points.
- Run a pilot, whether you build or buy a DToC, and compare results against a persona or customer-360 view over a statistically significant period using significant data. Ensure business and operating models can support the endeavor.
- Encourage customers to share data based on specific parameters. Define the DToC's benefits, agree to the level of control they will have over their data, including canceling the digital twin. Provide clear visibility into how their data will be used.

Sample Vendors

Absolutdata; edgeTI; Fetch.ai; Infogain; Nstream; Salesforce; ServiceNow; Tata Consultancy Services

Gartner Recommended Reading

A Digital Twin of a Customer Predicts the Best Customer Experience

Quick Answer: Privacy Basics for a Digital Twin of a Customer

Innovation Insight: Demystifying Digital Twin of a Customer for B2B Sales

Quick Answer: What Are the Capabilities of a Digital Twin of a Customer?

Supply Chain Executive Report: Drive Growth & Elevate Experiences With Digital Twin of the Customer

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 with the objective of sharing data and security insights universally. It enables 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 and evolving attack types, and reduce overhead caused by the proliferation and churn in security tool categories and attack types. Investing in composable, interoperable 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; this level of detail is crucial for 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.

- Organizations are looking for approaches such as CSMA to better integrate and interpret the outputs of siloed security technologies that operate with insufficient knowledge of other tools. Effective security and identity management requires a layered and integrated approach.

- 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 clarify and minimize the exposure that is consistent with business expectations.

Obstacles

- As vendors continue to support CSMA 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 costs.
- Organizations that choose to create their own CSMA construct will likely need significant engineering effort to integrate disparate products. Additionally, they might suffer if the security industry moves toward a set of interoperability standards after significant custom integration work has been completed.
- At the early stages of adoption, CSMA continues to evolve in response to consumer IT advancement and security technology consolidation. Planning for the relevant flexibility required to manage this change is difficult.
- Organizations understand and acknowledge the skills gaps and challenges in volumes of work but do not have clear solutions to deal with these issues.

User Recommendations

- Add purchasing requirements that focus on integration and interoperability of multivendor tools.
- Find your main security intelligence layer platform and connect the rest of the layers into it.
- Mature your security infrastructure by selecting product vendors who follow CSMA reference architecture, using standards such as open cybersecurity schema framework (OCSF), have fully developed advanced APIs, complete adherence to modern security standards, and integrations into security partner networks.
- Evolve your identity infrastructure into 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 for security teams.

Gartner Recommended Reading

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

2024 Planning Guide for Security

Emerging Technology Horizon for Information Security, 2022

How to Start Building a Cybersecurity Mesh Architecture

Disinformation Security

Analysis By: Dan Ayoub, Akif Khan

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Disinformation security is a suite of technologies that can address disinformation to help enterprises discern trust, protect their brand and secure their online presence. Technologies found within this category cover deepfake detection, impersonation prevention and reputation protection.

Why This Is Important

Disinformation can deceive or mislead an audience with fabricated information that leaves out critical context. Bad actors target organizations with disinformation content to manipulate buyers and commit fraud. Today, the availability of generative AI (GenAI) software capable of producing realistic and convincing deepfakes is at an all-time high. Disinformation security aims to address these threats and more by combining several technologies for methodological discernment of trust.

Business Impact

Disinformation can be weaponized to target an organization. Campaigns can quickly be put into action that will impact brand reputation, by degrading or distorting brand image, leading to loss of revenue. Targeted disinformation attacks have also been used to commit fraud against organizations. Disinformation impacts multiple business functions and teams, including executive leadership, security, public relations, marketing, finance and sales, which makes disinformation security an important part of an organization's security strategy.

Drivers

Disinformation security solutions are emerging in response to the growing threat landscape and incidence of disinformation attacks. For example:

- External threat actors target organizations with disinformation for ideological purposes, as part of cyber attacks and to commit fraud.
- Rapid improvement and advancements in adversarial AI increasingly make discerning legitimacy of multimedia content more difficult.
- Deepfake videos flood social media with disinformation. Large language models, coupled with GenAI voice cloning, can quickly create convincing disinformation content “spoken” in almost any voice.
- Security teams increasingly recognize that proactive defense necessitates looking beyond the organization’s public and private infrastructure. Many disinformation attacks are conducted using infrastructure not managed by the organization, from locations complicating proactive defense, removal, and legal team response.
- Marketing and public relations teams recognize the need to understand sentiment and any narratives that exist around an organization’s products, services, brands and executives.
- Organizations require digital watermarking and other proactive controls for content labeling systems to quickly and accurately identify legitimate, manipulated and GenAI-created content.
- Organizations need to consistently and reliably know who they’re dealing with online; this includes the ability to accurately identify candidates in the hiring process, employees, contractors, suppliers and customers.
- Organizations in regulated industries must comply with know-your-customer (KYC) requirements. The abuse of identity verification processes using GenAI-created content can expose organizations to serious regulatory sanctions.
- Account takeover is a major security concern for many organizations. Attack vectors are varied and can include using GenAI-created content to fool voice or face biometric authentication, or to trick identity verification processes used in account recovery workflows. Once authenticated as the perceived user, bad actors can take nefarious actions such as planting ransomware, stealing intellectual property, theft of funds and spreading disinformation.

Obstacles

- Many organizations are still coming to grips with understanding and quantifying the risks posed by disinformation. Developing proactive security plans is still an emerging trend within the market.

- Defensive technologies in this category are emerging; however, growing market hype can derail, confuse and complicate organizational strategies to combat disinformation.

- Emerging standards used to proactively label and identify GenAI-created content online, such as those from the Coalition for Content Provenance and Authenticity (C2PA), are still being finalized. Robust implementation that can be trusted will take time.

- New technologies are promising. However, maturation and effectiveness against sophisticated evolving adversaries remain to be seen. Many emerging solutions remain use-case-specific.

- Organizational awareness and responsibility for disinformation security is still low, further complicating and delaying responses in the event of a crisis.

- Some organizations still do not understand that proactive security necessitates looking beyond their own four walls to develop effective defensive strategies.

- Detecting the domains in which some disinformation attacks occur, such as real-time impersonation of an organization’s executive within a voice- or video-calling application, is a technical challenge, as many platforms are “walled garden” environments.

User Recommendations

- Adopt deepfake detection technologies that combine a variety of AI techniques and practices (including GenAI) and digital forensic capabilities to methodologically discern legitimate from artificial and fact from fiction.
- Implement impersonation prevention technologies to evaluate user behaviors holistically across all devices and interactions to validate authentic actions and provide nonrepudiation. Ensure that biometric systems involving face or voice have liveness detection capabilities deployed for employees, customers, suppliers and contractors.
- Create and carry out a proactive defense strategy. Protect the organization's reputation from external threat actors by proactively monitoring the systems, tools and infrastructure that adversaries may be using.
- Combat disinformation by forming cross-functional teams and strategies: Information security teams will continue to be responsible for management and administration of tools used to identify and detect disinformation impacting the organization. Public relations and marketing communications should handle external announcements and communications strategy.

Sample Vendors

Alethea; Blackbird Security; DuckDuckGoose; FaceTec; Hootsuite (Talkwalker); iProov; Reality Defender; Sensity; Sumsub

Gartner Recommended Reading

Emerging Tech: Combating Deepfakes, Deception and Disinformation in Multimedia Content

Market Guide for Identity Verification

Multiagent Systems

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

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

A multiagent system (MAS) is a type of AI system composed of multiple, independent (but interactive) agents, each capable of perceiving their environment and taking actions. Agents can be AI models, software programs, robots and other computational entities. Multiple agents can work toward a common goal that goes beyond the ability of individual agents, with increased adaptability and robustness.

Why This Is Important

Current AI is focused on the creation of individual agents built for specific use cases, limiting the potential business value of AI to simpler problems that can be solved by single monolithic models. The combined application of multiple autonomous agents can tackle complex tasks that individual agents cannot, while creating more adaptable, scalable and robust solutions. It is also able to succeed in environments where decentralized decision making is required.

Business Impact

Multiagent systems can be used in:

- **Generative AI:** Orchestrating AI agents for complex tasks
- **Robotics:** Swarms of robots and drones for warehouse optimization, search and rescue, environment monitoring, and other use cases
- **Energy and utilities:** Smart grid optimization and load balancing
- **Supply chain:** Optimizing scheduling, planning, routing and supply chain optimization
- **Telecom:** Network optimization and fault detection
- **Healthcare:** Using agents to model actors (individuals, households, professionals)

Drivers

- **Generative AI agents:** Large language models (LLMs) are increasingly augmented with additional capabilities, such as an internal memory and plug-ins to external applications, to implement AI agents. An emerging design pattern is to assemble and combine these LLM-based AI agents into more powerful systems, which is increasing the feasibility of and interest in multiagent systems.

- **Increased decision-making complexity:** AI is increasingly used in real-world engineering problems containing complex systems, where large networks of interacting parts exhibit emergent behavior that cannot be easily predicted. The decentralized nature of multiagent systems makes them more resilient and adaptable to complex decision making.

- **Simulation and multiagent reinforcement learning:** Advances in the realism and performance of simulation engines, as well as the use of new multiagent reinforcement learning techniques, allow for the training of multiagent AI systems in simulation environments, which can then be deployed in the real world.

Obstacles

- **Training complexity:** Multiagent systems are typically harder to train and build than individual AI agents. These systems can exhibit emergent behavior that is hard to predict in advance, which increases the need for robust training and testing.

- **Monitoring and governing multiple agents:** Coordination and collaboration between agents is challenging. Careful monitoring, governance and a common grounding are required to ensure that the combined multiagent system behavior achieves its intended goals.

- **Limited adoption and readiness:** Despite its benefits, the application of multiagent systems to real-world problems is not yet widespread, which creates a lack of enterprise awareness and readiness to implement.

- **Specialized skills required:** Building and deploying multiagent systems requires specialized skills beyond traditional AI skills, particularly the use of reinforcement learning and simulation.

- **Fragmented vendor landscape:** A fragmented vendor landscape inhibits customer adoption and engagement.

User Recommendations

- Use multiagent systems for complex problems that require decentralized decision making and cannot be solved by single AI agents. This includes problems with changing environments where agents need to adapt and problems where a diverse set of agents with different expertise can be combined to accomplish a goal.
- Shift to a multiagent approach gradually since this is an emerging area of research and the risks and benefits are not yet fully understood.
- Establish clear guardrails when implementing multiagent systems, including legal and ethical guidelines around autonomy, liability, robust security measures and data privacy protocols.
- Invest in the use of simulation technologies for AI training, as simulation is the primary environment to build and test multiagent systems.
- Educate your AI teams on multiagent systems, how they differ from single-agent AI design, and some of the available techniques to train and build these systems.

Sample Vendors

Alphabet; Ansys; Cosmo Tech; FLAME GPU; MathWorks; Microsoft; OpenAI; The AnyLogic Company

Gartner Recommended Reading

Innovation Insight: AI Agents

Innovation Insight: AI Simulation

AI Design Patterns for Large Language Models

6G

Analysis By: Kosei Takiishi

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

6G is the upcoming cellular network technology that will follow 5G. In 2024, the features and timetable for 6G are not clearly defined, although it's expected to be commercialized around 2028 by some communications service provider (CSP) pioneers. 6G will enhance 5G capabilities and is intended to provide higher peak data rate (such as 100 Gbps to 1 Tbps), lower latency (such as 0.1 ms) and much more connection density and energy efficiency (potentially 10 times more efficient).

Why This Is Important

Within four years, CSPs will start deploying 6G networks, driven by governments and market forces to demonstrate technical leadership and network infrastructure optimization. The U.N.'s Agenda for Sustainable Development sets ambitious goals for 2030, and 6G is expected to address some of these. Design and research for 6G is already underway by many industrial associations and academic and commercial organizations.

Business Impact

6G will enable end users, including consumers and enterprises, to transfer and process large volumes of data in real time, which enables true immersive experiences and more mission-critical human-machine communications, based on improved security, energy efficiency, nonterrestrial networks and AI. Much richer and advanced connectivity of the physical world with the digital world — digital-physical fusion — is expected. 6G will support use cases such as immersive communication and coverage in unmanned areas.

Drivers

- Government and related funding are driving a lot of research. In February 2023, the South Korean Minister for Science and ICT unveiled the K-Network 2030 plan, calling for South Korean tech firms to lead the way in developing world-class 6G technologies and software-based networks. In November 2023, the European Smart Networks and Services Joint Undertaking announced the new Research and Innovation work program for 2024, including information on Horizon Europe's 2024 Call for Proposals. India launched the Bharat 6G Alliance to drive innovation and collaboration in July 2023. Under India's Digital Communication Innovation Square, a grant of Rs. 48 crore aided 66 startups. There is also a grant of Rs. 240.5 crores under the Telecom Technology Development Fund scheme.

- More academics and commercial organizations want to be part of the 6G process, and active research has already begun. China Unicom published a 6G network architecture white paper in June 2023. The National Institute of Information and Communication Technology in Japan updated its Beyond 5G/6G White Paper version 3.0 in June 2023. SK Telecom in South Korea released a white paper, "5G Lesson Learned, 6G Key Requirements, 6G Network Evolution, and 6G Spectrum," in August 2023.

- Many commercial organizations and academic institutions have started their 6G research to be a part of the future 6G patent pool.

- Adoption of new releases of 3rd Generation Partnership Project (3GPP) standards is broadly inevitable because it eventually happens as a result of network and endpoint upgrades.

- AI and machine learning are expected to be deeply integrated into 6G networks. This integration will enable smarter, more efficient network operations and management, and support advanced AI-driven applications adoption.

Obstacles

- Five years have passed since 5G commercialization in developed countries, but its monetization and use cases are not clear. Success or failure of 5G to drive revenue and new business opportunities will have a major impact on the timing of 6G commercialization and business.
- The telecommunications industry has formulated its own specifications and standardization (such as 2G, 3G, 4G and 5G). It is unclear whether 6G will be able to incorporate external opinions, extending the start provided by some other industries' participation in developing 5G standards.
- A bifurcated geopolitical environment will challenge the creation of global standards, with countries (and companies) vying to include their patented technology in the new standard.
- Some 6G technologies, such as THz wireless, may not prove to be technically viable or cost-effective for most cellular users' needs.

User Recommendations

- Monitor discussion of the currently emerging 6G carefully by tracking 3GPP working groups, requirement studies and emerging use cases.
- Prepare early trials and proofs of concept in the late 2020s with vendors to learn more about the capabilities of 6G, the required infrastructure and operation changes, and early use cases. Then, begin building skill sets.
- Support your regulators and government to create their new national policy for 5G-Advanced and 6G. Technology innovation and strategy leaders should look at evolving 6G standards to get an early idea of future networking technologies.

Sample Vendors

Ericsson; Huawei Technologies; Nokia; NTT DOCOMO; Qualcomm; Samsung Electronics; SK Telecom

Gartner Recommended Reading

Emerging Tech Impact Radar: Communications

AI Supercomputing

Analysis By: Chirag Dekate

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

AI supercomputing is a rapidly evolving domain where purpose-designed supercomputing systems combine state-of-the-art innovations in computational accelerators, specialized software, high-speed networks and performance-optimized storage to create an integrated platform that accelerates training and deployment of complex, computationally intensive AI models such as generative AI (GenAI) models.

Why This Is Important

Purpose-designed AI supercomputers offering extreme scale parallelism are crucial to driving innovation in advanced GenAI, data analytics and simulation environments. AI supercomputing is important because it enables processing of extreme scale datasets that are crucial to developing and applying innovative GenAI models. As GenAI evolves beyond large language models (LLMs) into more complex domains including multimodality, the computational capability required can only be addressed using advanced AI supercomputers.

Business Impact

AI supercomputing transforms businesses by:

- Powering enterprises to deliver game-changing AI through customization of GenAI models to their organizational data and application context.
- Enabling AI model innovators to develop and deliver new breakthrough generative AI models including domain-specific models, different-sized GenAI models and multimodal models.
- Accelerating large-scale inference of GenAI models and driving value creation in both everyday AI and game-changing AI.

Drivers

- **Massive datasets:** Leadership-class GenAI models rely on ingestion of internet-scale datasets for training. As the model mix evolves beyond LLMs to multimodal models, the resultant growth of data (both structured and unstructured) will drive the need for advanced AI supercomputing resources.

- **Algorithmic complexity:** GenAI today comprises a broad mix of different-sized models, domain-specific models and multimodal models. All of these models rely on extreme scale parallelism offered by advanced AI supercomputers for both training and inference.

- **Accelerated adoption of GenAI:** Practical applications of GenAI across industries, especially game-changing AI, will require speed and accuracy that only AI supercomputers can provide.

- **Regional regulatory requirements:** Most of the early AI supercomputing capacity is clustered in the U.S. and Western Europe. The combinatorial impact of data residency requirements and the accelerated pace of global adoption of GenAI will drive the need for the delivery of AI supercomputing everywhere.

- **Competition and investment:** Extreme hype and the promise of transformative impact is driving intense competition globally to create new, innovative generative AI models that broaden practical applicability of GenAI. Cloud service providers, hyperscalers and specialized hosting and colocation providers are investing billions of dollars in building advanced AI supercomputing foundations to address this demand.

Obstacles

- **Supply chain risks:** More than 92% of advanced semiconductor manufacturing that powers AI supercomputers is sourced from TSMC. Geopolitical disruptions or natural disasters (such as earthquakes and typhoons) that disrupt AI supercomputing supply chains could pose significant risks for the market.

- **Energy constraints:** Graphics processing unit (GPU)-based AI supercomputing technologies require multi-MW scale datacenters that enterprises may not be able to address due to limitations of their existing facilities or their local utility provider.

- **Component challenges:** GPU-based AI supercomputing technologies integrate complex sets of components from high-bandwidth memory stacks to transformers that downshift power to GPU cores. Supply chain disruptions in these components will create challenges for growth of AI supercomputing.

- **Lack of value delivery from GenAI:** If the extreme hype in GenAI doesn't translate to broad-based value creation, it could pose obstacles to continued progress of AI supercomputing.

User Recommendations

- **Choose cloud over on-premises:** Cloud-based AI supercomputing offers access to innovations beyond GPUs. Utilize them where possible to deliver better economies of scale. Use cloud-based resources when possible. When no alternatives exist, devise on-premises AI supercomputing strategies.

- **Adopt use-case-driven AI supercomputing:** Focus on identifying use cases that demand extreme computational power, such as GenAI training, retrieval-augmented generation, fine-tuning, and inference and deployment of GenAI models. Use cloud-based AI supercomputing for scoping AI supercomputing resources to derisk and rightsize your investments.

- **Select the right architecture:** AI supercomputers based on GPUs and the alternative AI application-specific integrated circuits available via leading cloud providers and innovative AI supercomputing vendors can deliver better performance, accuracy and scalability. Not all models require acceleration. The latest CPUs with embedded AI acceleration can help train and infer models that are below 7B parameters.

Sample Vendors

Amazon; AMD; Google; Groq; IBM; Intel; Meta; Microsoft; NVIDIA; SambaNova Systems

Gartner Recommended Reading

A Comparison of Generative AI Platform Offerings

Market Guide for Specialty Cloud Providers

Autonomous Agents

Analysis By: Christian Stephan

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Embryonic

Definition:

Autonomous agents are combined systems that achieve defined goals without human intervention. They use a variety of AI techniques to identify patterns in their environment, make decisions, execute a sequence of actions and generate outputs. These agents have the potential to learn from their environment and improve over time, enabling them to handle complex tasks.

Why This Is Important

Autonomous agents represent a significant shift in AI capabilities. By their independent operation and decision capabilities, they can improve business operations, enhance customer experiences and enable new products and services. On the other hand, tech executives need to manage new challenges in transparency, ethics and workforce adoption. The early stage is inflating expectations, despite most agents providing limited use, but the fast development in this area demands proper observation.

Business Impact

Business operations could be significantly enhanced with autonomous agents by boosting efficiency through automating complex tasks, which improves organizational productivity. They might also enhance customer experience with 24/7 personalized service via conversational agents empowered with robotic process automation capabilities to trigger organization processes. This will probably come with cost savings, granting a competitive edge. It also poses an organizational shift of workforce from delivery to supervision.

Drivers

- **Investment and progress in AI development:** Rapid advancements in AI, particularly in the capability of large language models (LLMs) and reinforcement learning, are favoring autonomous agent development.

- **Complexity of IT landscape:** The ability to decompose high-level tasks into processes coupled with multimodal inputs/outputs allows autonomous agents to interpret their environment and handle the plurality of systems.

- **Transfer learning and adaptive AI:** Applying learned knowledge to new tasks or adapting to changing environments will enhance the versatility and effectiveness of autonomous agents.

- **Multimodality:** The ability to process and integrate multiple types of data (such as text, images and audio) enables autonomous agents to handle more complex tasks and to engage with software interfaces to take action.

- **Digitalization of organizations:** By digitizing assets and processes for effective operations, companies have generated data and automation that autonomous agents can connect to in order to perform complex tasks in the environment.

- **Demand for personalization:** In an era of increasing customer expectations on personalized experiences and high availability, autonomous agents can deliver such extended services at scale.

- **Promise on cost savings:** Setting up an autonomous infrastructure always bears opportunity for savings on human labor that could probably be replaced by IT infrastructure.

- **Increasing vendor offerings:** Large vendors are explicitly marketing their offerings of early agents and assistants.

Obstacles

- **AI trust, risk and security management challenges:** Autonomous agents most likely will have access to sensitive information and critical infrastructure and need effective protection. Transparent and explainable decision making is mandatory.
- **Lack of human supervision:** Without a human in the loop, the potential to mitigate AI errors is reduced.
- **Upcoming regulations:** Autonomy, in particular, is a much-discussed and debated policy issue. Early regulatory proposals point to strict regulations and liabilities for autonomous actors.
- **Organizational resistance:** The fear of replacement will expose autonomous agents to strong resistance within the workforce.
- **Decision making and liability:** When AI agents gain more autonomy and handle more complex tasks, the sometimes-unreliable output of generative AI models can steer organizations into faulty or even illegal decisions.

User Recommendations

- **Grow experience:** Invest in understanding and integrating AI agents in your strategic planning, considering their growing autonomy and wide-ranging applicability.
- **Safeguard the operation:** Define clear process guardrails, including legal and ethical guidelines around autonomy, liability, robust security measures and data privacy protocols.
- **Observe and experiment:** The technology is immature and needs to evolve. Explore the potential of multiagent systems capable of operating both collaboratively and independently to enhance flexibility and adaptability in response to various tasks and scenarios.
- **Develop a data strategy:** Autonomous agents will require high-quality data to function effectively. Develop and implement strategies for data collection, cleaning, management and privacy for your organization.
- **Invest in AI and data literacy:** Invest in training your workforce. This will help to understand the potential and limitations of autonomous agents and how to work with them effectively.

Sample Vendors

Accelint (Soar Technology); Autochain; Cognosys; Dropzone AI; Dust; Fluid AI; LangChain; LlamaIndex; MultiOn; NinjaTech AI

Gartner Recommended Reading

Innovation Insight: AI Agents

Autonomous Things: Technology Use Cases for R&D

Emerging Tech: Tech Innovators in Tabular Synthetic Data — Domain-Focused

2023 Utility Trend: Establish Decision Intelligence Before Chasing Autonomous Business

Digital Immune System

Analysis By: Joachim Herschmann, Jim Scheibmeir

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Definition:

The digital immune system (DIS) approach to building applications interlinks practices from six areas: observability, software testing, chaos engineering, automated remediation, site reliability engineering (SRE) and software supply chain security. This approach ensures better quality of applications. It focuses on making applications more resilient so they recover quickly from failures, rather than impacting users and business performance.

Why This Is Important

Software development teams are under pressure to deliver faster, while dealing with increasingly complex architectures, compliance needs and constantly evolving technology stacks. This exposes organizations to operational and business risks when applications and services are either severely compromised or stop working altogether. The DIS approach combines and amplifies several resilience engineering practices to mitigate these risks and deliver greater business impact.

Business Impact

A DIS strategy helps engineering teams build and deploy software confidently, as they can continuously assess system health, remediate issues and optimize system performance. Also, systems designed for digital immunity help identify code quality issues related to performance, security and other operational attributes early in the software development life cycle. The DIS approach creates software engineering teams focused on connecting software development to business value.

Drivers

DIS adoption will grow over time, as this approach to building applications enables:

- **Troubleshooting in highly distributed, cloud-based applications:** Software engineering leaders struggle to plan for all eventualities of how modern, highly distributed software systems may fail. Troubleshooting issues in distributed applications requires fundamentally different techniques, compared with monolithic or client/server applications.

- **Business continuity:** As an enabler of business operations, IT plays a significant role in business continuity and enterprise success. In response to pressure from business stakeholders, software engineering leaders are looking for new ways to improve the resilience of business-critical systems and reduce their vulnerability.

- **Risk mitigation:** Prominent examples of how leading digital enterprises have suffered from software glitches, outages or security incidents put additional pressure on software engineering leaders to mitigate risks to the business.

- **Continuous improvement:** Organizations need a frame of reference to improve the resilience of digital systems, measure progress and use failures as learning opportunities. Teams must continuously improve their skills to minimize the impact of a failure, with a focus on preserving the overall health of the system.

- **Better developer experience:** Developers often get disrupted in their daily work, as they have to deal with incidents and frequent firefighting exercises, which are detrimental to developer experience. A DIS approach fosters a mindset of continuous quality improvement, and enables a shift toward building quality and security into the product.

- **Reliability as a key differentiator:** Organizations that have recognized reliability as a differentiator in the market need key insights, operational agility and competence. They acquire these through the combination of observability, AI-augmented testing, chaos engineering, SRE and security practices.

Obstacles

- **Technology focus:** Tools are enablers of a digital immune system strategy, but just throwing technology at a problem will not solve it. Tools automate and facilitate processes and practices, but the processes and practices need to be in place first and evolve over time.

- **Failure to measure:** The core elements of a digital immune system are not new. But most organizations are only using some parts of it, often in uncoordinated ways, and without a plan to close gaps and evolve them into a more holistic strategy. It is, therefore, important to measure progress in the evolution of the different practices and the success of technologies that enable them.

- **Internal pushback:** Implementing a digital immune system requires engaging stakeholders across the organization and seeking out opportunities for collaboration and improvement. Such a holistic approach can be seen as too far-reaching and lead to turf wars.

User Recommendations

- Adopt a DIS strategy as a standard software engineering approach to prepare your teams to handle unexpected and unforeseeable system behaviors, quickly remediate software defects, and minimize the impact on users. Support a “you build it, you run it” approach.

- Assess which business capabilities have the highest priority or will benefit the most from DIS investments by analyzing current gaps in software quality.

- Establish a platform engineering team to build standard platform support for a DIS. Create dedicated communities of practice (CoPs) to share lessons learned, guiding principles, reusable assets, standards, tools and any AI-based insights realized.

- Foster collaboration on DIS opportunities by encouraging and rewarding resilience improvements across the organization.

- Foster a collaborative culture between development, security and operations teams to ensure ongoing support for DIS initiatives.

Sample Vendors

Antithesis; Dynatrace; Harness

Gartner Recommended Reading

Improve Software Quality by Building Digital Immunity

Innovation Insight: Continuous Quality

Market Guide for AI-Augmented Software-Testing Tools

Artificial General Intelligence

Analysis By: Pieter den Hamer

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Definition:

Artificial general intelligence (AGI), also known as strong AI, is the (currently hypothetical) intelligence of a machine that can accomplish any intellectual task that a human can perform. AGI is a trait attributed to future autonomous AI systems that can achieve goals in a wide range of real or virtual environments at least as effectively as humans can.

Why This Is Important

As AI becomes more sophisticated and powerful, with recent great advances in generative AI (GenAI) in particular, a growing number of people see AGI as no longer purely hypothetical. Improving the understanding of at least the concept of AGI is critical for steering and regulating AI's further evolution. It is also important to manage realistic expectations and to avoid prematurely anthropomorphizing AI. However, if AGI becomes real, its impact on the economy, (geo)politics, culture and society cannot be underestimated.

Business Impact

In the short term, organizations must know that the hype about AGI exists today among many stakeholders, stoking fears and unrealistic expectations about current AI's true capabilities. This AGI anticipation is already accelerating the emergence of more AI regulations and affects people's trust and willingness to apply AI today. In the long term, AI continues to grow in power and, with or without AGI, will increasingly impact organizations, including the advent of machine customers and autonomous business.

Drivers

- Recent great advances in applications of GenAI and the use of foundation models and large language or multimodal models drive considerable hype about AGI. These advances have been enabled largely by the massive scaling of deep learning, as well as by the availability of huge amounts of data and compute power. To further evolve AI toward AGI, however, current AI will need to be complemented by other (partially new) approaches, such as knowledge graphs, multiagent systems, simulations, evolutionary algorithms, causal AI, composite AI and likely other innovations yet unknown.

- Vendors such as Google, IBM and OpenAI are openly discussing and actively researching the field of AGI, creating the impression that AGI lies within reach. However, their definitions of AGI vary greatly and are often open to multiple interpretations.

- Humans' innate desire to set lofty goals is also a major driver for AGI. At one point in history, humans wanted to fly by mimicking bird flight. Today, airplane travel is a reality. The inquisitiveness of the human mind, taking inspiration from nature and from itself, is not going to fizzle out.

- People's tendency to anthropomorphize nonliving entities also applies to AI-powered machines. This has been fueled by the humanlike responses of ChatGPT and similar AI, as well as AI being able to pass several higher-level education exams.

- Complex AI systems display behavior that has not been explicitly programmed. Among other reasons, this results from the dynamic interactions between many system components. Consequently, AI is increasingly attributed with humanlike characteristics, such as understanding. Although many philosophers, neuropsychologists and other scientists consider this attribution as going too far or being highly uncertain, it has created a sense that AGI is within reach or at least is getting closer. In turn, this has triggered massive media attention, several calls for regulation to manage the risks of AGI and a great appetite to invest in AI for economic, societal and geopolitical reasons.

Obstacles

- Unreliability, lack of transparency and limited reasoning capabilities in current AI are not easy to overcome with the intrinsically probabilistic approach of deep learning. More data or more compute power for ever-bigger models are unlikely to resolve these issues, let alone to achieve AGI.
- The meanings of "intelligence" and related terminology like "understanding" have little scientific consensus, including the definition and interpretation of AGI. Scientific understanding about human intelligence is still challenged by the enormous complexity of the human brain and mind. Any claims about AGI — in whatever form it may emerge — are hard to validate when we don't even understand human intelligence. However, even when AGI will be properly understood and defined, further technological innovations will likely be needed to implement AGI. Therefore, AGI as defined here is unlikely to emerge in the near future.
- If AGI materializes, it is likely to lead to the emergence of autonomous actors that, in time, will be attributed with full self-learning, agency, identity and perhaps even morality. This will open up a bevy of legal rights of AI and trigger profound ethical and even religious discussions. AGI also brings the risk of negative impacts on humans, from job losses to a new, AI-triggered arms race and more. This may lead to serious backlash, and regulations to ban or control AGI are likely to emerge in the near future.

User Recommendations

- Engage with stakeholders to address their concerns and create or maintain realistic expectations. Today, people may be either overly concerned about future AI replacing humanity or overly excited about current AI's capabilities and impact on business. Both cases hamper a realistic and effective approach to using AI today.
- Stay apprised of scientific and innovative breakthroughs that may indicate the possible emergence of AGI. Meanwhile, keep applying current AI to learn, reap its benefits and develop practices for its responsible use.
- Although AGI is not a reality now, current AI already poses significant risks regarding bias, reliability and other areas. Prepare for emerging AI regulations and promote internal AI governance to manage current and emerging future risks of AI.

Sample Vendors

Aigo; Google; IBM; Microsoft; OpenAI

Gartner Recommended Reading

The Future of AI: Reshaping Society

Innovation Insight for Generative AI

Innovation Insight: AI Simulation

Applying AI — Key Trends and Futures

Innovation Insight for Artificial Intelligence Foundation Models

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) in which 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 to maximize the overall reward while minimizing punishments.

Why This Is Important

When other ML approaches are not feasible due to a lack of labeled training data, some problems can best be solved with RL. The technology can lead to significant improvements in self-driving cars, robotics, vehicle routing, warehouse optimization, logistics, predictive maintenance and other industrial control scenarios. Like many new and longstanding AI techniques, reinforcement learning is primed to undergo a period of mass experimentation.

Business Impact

RL's near-term potential is greatest in industrial control and design, marketing and advertising, recommendation systems and gaming industries. As simulation, synthetic data and agent-based modeling gain traction, reinforcement learning will become increasingly common in business tools. AI and data science teams of all sizes are adding reinforcement learning to their portfolio of techniques.

Drivers

- Recent successes across various industries (for example, text summarization and machine translation, real-time bidding for marketing and advertising, creating 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 that enable more application scenarios for RL
- Better simulation capability
- 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 designing the RL model's reward structure for most business scenarios
- Often brittle or difficult-to-implement solutions that apply only in limited use cases
- Lack of staff with reinforcement learning experience
- Lack of explainability

User Recommendations

- Apply RL in use cases that require frequent model retraining with traditional techniques, since RL can adapt to new environments 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; Dataiku; MathWorks; Microsoft; TensorFlow

Gartner Recommended Reading

Innovation Insight: AI Agents

Innovation Insight: AI Simulation

Go Beyond Machine Learning and Leverage Other AI Approaches

Federated Machine Learning

Analysis By: Tong Zhang, Svetlana Sicular, Mike Fang, Ben Yan, Bart Willemsen

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

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

Why This Is Important

FedML highlights an important innovation in (re)training ML algorithms in a decentralized environment without disclosing sensitive business information. It enables more personalized experiences with local learning in smartphones, softbots, autonomous vehicles or Internet of Things (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 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 healthcare quality improvement of 16% and a 38% increase in generalization over local models, as a collaboration result of 20 institutes. For more information on this report, see Federated Learning for Predicting Clinical Outcomes in Patients With COVID-19, Nature Medicine.

Drivers

- The proliferation of privacy and legislative regulations require protection of local data. FedML enables the protection of data privacy.
- As large language model (LLM) evolves, research on federated LLM emerges so that a group of organizations could collaborate to train LLMs together.
- 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 operational efficiency.
- 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

- FedML is still not widely known in the enterprise, as it lacks marketing on the vendor and researcher sides.
- 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.
- 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 headstart to decentral ML applications by deploying a common, centrally pretrained model, while still providing personalization and contextualization by locally retraining the model based on 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; Eder Labs; FedML; Google; Intel; NVIDIA; Owkin; WeBank; WithSecure

At the Peak

AI TRiSM

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

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

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

Why This Is Important

AI models and applications should be subject to protection mechanisms during development and at runtime. 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, privacy and security controls and trust enablers that helps enterprises govern and manage AI models and applications' life cycles — and accomplish business goals. The benefit is improved outcomes and performance and enhanced compliance with regulations such as 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, AI misperformance and compromised data confidentiality. Inaccurate, unethical or unintended AI outcomes, process errors, uncontrolled biases, and interference from benign or 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 or incorrect business decisions.

Drivers

- OpenAI's ChatGPT democratized third-party generative AI applications 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 AI applications often pose considerable data confidentiality risks. This is partly because large, sensitive datasets used to train AI models often come from various sources, including data shared by users of these applications.

- 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 adjust bias controls where needed in both (training) data and algorithmic functions.

- AI outputs that are unchecked can steer organizations into faulty decision making or harmful acts because of inaccurate, illegal or fictional information driving business decisions.

- AI model and application explainability and expected behavior must be constantly tested through observation and testing of model and application outputs. Doing so ensures original explanations, interpretations and expectations of AI models and applications remain active during model and application 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 and application risk. Such regulations define the 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.
- Off-the-shelf software that embeds AI is often closed and does not support APIs to third-party products that can enforce enterprise policies.
- 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 unit to manage AI TRiSM. Include members with a vested interest in AI projects.
- Define acceptable use policies at a level granular enough to enforce.
- Implement data classification and permissioning systems to ensure enterprise policies can be enforced.
- Establish a system to record and approve all AI-based applications and gain periodic user attestation that they are used according to preset intentions.
- Use appropriate AI TRiSM toolsets to manage AI model, application, and agent trust risk and security.
- Require vendors with AI components to provide verifiable attestations of expected AI behavior.

- Implement AI data protection solutions and use different methods for different use cases and components.
- Establish data protection and privacy assurances in license agreements with vendors hosting LLMs.
- Constantly validate and test the security, safety and risk posture of all AI used in your organization, no matter the footprint.

Sample Vendors

Aporia; Bosch Global Software Technologies (AIShield); Harmonic; Lasso Security; ModelOp; Prompt Security; Protopia AI; TrojAI

Gartner Recommended Reading

Top Strategic Technology Trends for 2024: AI Trust, Risk and Security Management

Innovation Guide for Generative AI in Trust, Risk and Security Management

Market Guide for AI Trust, Risk and Security Management

Tool: Generative AI Policy Kit

Quick Answer: The EU AI Act and Its Anticipated Impact

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 cryptographic 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, there is usually a performance trade-off.

Why This Is Important

HE offers an 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 (secure multiparty computation [SMPC]).
- Share all or part of users' data, while protecting their privacy.
- Base its system on lattice encryption, which is a postquantum cryptography (PQC) algorithm.

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 SMPC will benefit internal and external protection of data.

- With the adoption of new NIST PQC algorithms, the homomorphic properties of lattice are being explored by more clients. Timely adoption of HE in data protection will sustainably protect data against the threat of "harvest now, decrypt later" attacks.

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.

- 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.

- 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.

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 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 Cybersecurity; Ziroh Labs

Gartner Recommended Reading

Emerging Technologies: Homomorphic Encryption for Data Sharing With Privacy

Emerging Tech Impact Radar: Security

Three Critical Use Cases for Privacy-Enhancing Computation Techniques

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

Prompt Engineering

Analysis By: Frances Karamouzis, Jim Hare, Afraz Jaffri

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Definition:

Prompt engineering is the discipline of providing inputs, in the form of text or images, to generative AI (GenAI) models to specify and confine the set of responses the model can produce. The inputs prompt a set that produces a desired outcome without updating the actual weights of the model (as done with fine-tuning). Prompt engineering is also referred to as “in-context learning,” where examples are provided to further guide the model.

Why This Is Important

Prompt engineering is the linchpin to business alignment for desired outcomes. It is important because large language models (LLMs) and GenAI models in general are extremely sensitive to nuances and small variations in input. A slight tweak can change an incorrect answer to one that is usable as an output. Each model has its own sensitivity level, and the discipline of prompt engineering is to uncover the sensitivity through iterative testing and evaluation.

Business Impact

Prompt engineering has the following business impacts:

- **Performance:** It helps improve model performance and reduce hallucinations.
- **Business alignment**: It allows subject data scientists, subject matter experts and software engineers to steer foundation models, which are general-purpose in nature, to align to the business, domain and industry.
- **Time to market, quality, efficiency and effectiveness:** There are a number of architecture options as well as execution options that AI leaders must balance. There is also a myriad of prompt optimization tools that will diminish (or at the very least shift) the need for manual engineering.

Drivers

- **Balance and efficiency:** The fundamental driver for prompt engineering is it allows organizations to strike a balance between consuming an “as is” offering versus pursuing a more expensive and time-consuming approach of fine-tuning. GenAI models, and in particular LLMs, are pretrained, so the data that enterprises want to use with these models cannot be added to the training set. Instead, prompts can be used to feed content to the model with an instruction to carry out a function.

- **Process or task-specific customizations or new use cases:** The insertion of context and patterns that a model uses to influence the output generated allows for customizations for a particular enterprise or domain, or regulatory items. Prompts are created to help improve the quality for different use cases — such as domain-specific question answering, summarization, categorization, and so on — with or without the need for fine-tuning a model, which can be expensive or impractical. This would also apply to creating and designing new use cases that utilize the model’s capability for image and text generation.

- **Validation and verification:** It is important to test, understand and document the limits and weaknesses of the models to ensure a reduced risk of hallucination and unwanted outputs.

Obstacles

- **Prompt engineering is a new discipline:** The craft of designing and optimizing user requests to an LLM or LLM-based chatbot to get the most effective result is still emerging. Engineers are finding that desired outputs using GenAI can be challenging to create, debug, validate and repeat. Communities worldwide are developing new prompt engineering methods and techniques to help achieve these desirable outcomes.

- **Approaches, techniques and scalability:** A unified approach to performing prompt engineering does not exist. Complex scenarios need to be broken down into smaller elements. It is challenging to debug complex prompts. Understanding how specific prompt elements influence the logic of the LLM is vital. Scalable and maintainable methods of prompt engineering are still a work in progress for most organizations.

- **Role alignment:** Data scientists are critical to understanding the capabilities and limits of models, and to determining whether to pursue a purely prompt-based or fine-tuning-based approach (or combination of approaches) for customization. The ultimate goal is to use machine learning itself to generate the best prompts and achieve automated prompt optimization. This is in contrast to an end user of an LLM who concentrates on prompt design to manually alter prompts to give better responses.

User Recommendations

- Build awareness and understanding of prompt engineering to quickly start the journey of shape-shifting the appropriate prompt engineering discipline and teams.

- Build critical skills among different team members that will synergistically contribute critical elements. For example, there are important roles for data scientists, business users, domain experts, software engineers and citizen developers.

- Educate the team with the myriad of options of prompt optimization tools that will diminish (or at the very least shift) the need for manual engineering.

- Communicate and cascade the message that prompt engineering is not foolproof. Enterprise teams apply rigor and diligence to permeate and work to ensure successful solutions.

Sample Vendors

FlowGPT; Google; HoneyHive; Magniv; Microsoft; PromptBase; Salesforce

Gartner Recommended Reading

How to Engineer Effective Prompts for Large Language Models

Prompt Engineering With Enterprise Information for LLMs and GenAI

Quick Answer: How Will Prompt Engineering Impact the Work of Data Scientists?

Generative AI Changes Software Engineering Leaders' Responsibilities

GitOps

Analysis By: Paul Delory, Arun Chandrasekaran

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

GitOps is a type of closed-loop control system for cloud-native applications. The term is often used more expansively, usually as shorthand for automated operations or continuous integration/continuous deployment (CI/CD), but this is incorrect. According to the canonical OpenGitOps standard, the state of any system managed by GitOps must be expressed declaratively, versioned and immutable, pulled automatically, and continuously reconciled.

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, 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, GitOps can, 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 many 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; however, 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 Argo CD.

- **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 buy. It is 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

Akuity; GitLab; Harness; Red Hat; Upbound

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

Machine Customers

Analysis By: Don Scheibenreif, Mark Raskino

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Definition:

Machine customers are nonhuman economic actors that obtain goods or services in exchange for payment. Examples of machine customers include virtual personal assistants, smart appliances, connected cars and Internet of Things (IoT)-enabled factory equipment. Machine customers act on behalf of a human customer or an organization.

Why This Is Important

Currently, there are 15 billion internet-connected machines with the potential to act as customers. We expect the number of machine customers, such as virtual assistants with AI capabilities, to rise steadily over time. Machines are increasingly gaining the capacity to buy, sell and request services. Further, machine customers will advance beyond the role of simple informers to advisors and, ultimately, decision makers.

Business Impact

Over time, trillions of dollars are expected to be in control of nonhuman customers. This will result in new opportunities for revenue, efficiencies and managing customer relationships. Leaders seeking new growth must reimagine their operating and business models to take advantage of this emerging market of tens of billions of machine customers. Organizations that miss this opportunity will be marginalized, just like those retailers who missed the digital commerce wave.

Drivers

- In the next few years, machine customers are expected to become significant players in the industrial manufacturing, retail and consumer industries.
- In the forthcoming years, billions of connected products will have the potential to behave as customers — that is, to shop for services and supplies for themselves and their owners.
- Currently, most machines merely inform or make simple recommendations. However, some machines are emerging as more complex customers. For example, HP Instant Ink is a service that enables connected printers to automatically order their own ink when supplies run low. Also, some Tesla cars already order their own spare parts, and Amazon offers its Dash Replenishment Service for a variety of household appliances. Advances in generative AI, and applications like ChatGPT, Rabbit Inc.'s R1 device, and Amazon's Rufus, will accelerate the development and deployment of machine customers. These tools can diagnose and break down complex tasks to make the right recommendations, service requests and other functions.
- Technology advancement allows machines to take over more complex tasks that required human efforts in the past. In B2B, Pactum AI is an AI-based contract negotiating system. It evaluates agreement terms and offers an unbiased resolution that can result in either business development or renewed agreements where both parties have been equally evaluated to support a fair deal. Based on historical data, terms are suggested and discussed through an AI-based chatbot interface to generate a contract that is then ready for review and signing. Walmart is using them to negotiate with suppliers of goods not for resale.
- Machine customers have the potential to generate new revenue opportunities, increase productivity and efficiency, improve health and well-being, and enhance the security of physical assets and people. They will benefit both selling and buying organizations.

Obstacles

- **Operating models:** Marketing, selling and serving a machine customer will upend your operating model. A new definition of customer experience for a machine customer will be needed.
- **Lack of trust:** Humans may not trust the machine customer technology they use to predict, execute and maintain privacy accurately. Conversely, machine customers may not trust the supplier organization to do the same.
- **Fear of machines:** Some humans may initially be uneasy about delegating purchasing functions to machines. Organizations must consider what ethical standards, legal compliance, fraud and risk mitigation are needed to operate in a world of machine customers.

User Recommendations

- Identify specific use cases where your products and services can be extended to machine customers. Initiate collaboration with your chief digital officer, chief data officer, chief strategy officer, sales leaders and chief customer officer to explore the business potential of machine customers.
- Assess B2B customers' tech purchase intent data to identify the machine customer capabilities and use cases customers are deploying.
- Pilot the ideas compiled during the identification of use cases to understand the technologies, processes and skills required to implement machine customers adequately.
- Build your organization's capabilities around digital commerce and AI, especially generative AI, for the next few years. Use APIs and enterprise bots to enable machine customers for low-complexity transactions, such as autoreplenishment. Then, extend your capabilities for more complex purchases.
- Follow examples from organizations such as Amazon, iProd, OpenAI, NTT Data, Google, HP Inc. and Tesla for evidence of capabilities and business-model impact.

Sample Vendors

Adept; Amazon; Google; HP Inc.; iProd; NTT DATA; OpenAI; Pactum AI; rabbit; Tesla

Gartner Recommended Reading

Top Strategic Technology Trends for 2024: Machine Customers

Machine Customers Are Coming to Disrupt B2B Sales

Podcast: When Machines Become Customers

A New Era in Digital Banking: How CIOs Can Prepare for Machine Customers

Quick Answer: The Skills Your Team Needs to Compete in the Machine Customer Market

Internal Developer Portals

Analysis By: Manjunath Bhat

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Internal developer portals enable developers to discover, access and operate internal developer platform components, tools, services and knowledge assets. This helps improve developer experience and service reliability while ensuring centralized governance and shared visibility across multiple teams. Capabilities include service and resource catalogs, software quality and security scorecards, scaffolding templates for building new components and plug-ins for integrating with other dependencies.

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 and automation of 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 experience and productivity —** Helps development teams improve their delivery cadence by improving developer experience, reducing cognitive load and shortening the feedback loop.

- **Reliability and resilience —** Aims to provide visibility to application health and include scorecards to objectively assess their production readiness.

- **Security and governance —** Includes 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 for multiple development teams. Platform teams curate internal developer platforms to rationalize and unify the siloed systems and processes. Internal developer portals serve as the user 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 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 rein in complexity will help reduce cognitive load for developers.

- **Innersource —** To enable rapid innovation and facilitate greater collaboration and knowledge sharing, software engineering leaders are adopting innersource approaches to software development. Fannie Mae built an innersource portal to provide teams with visibility into what their peers are working on and drive the standardization and community engagement necessary to a culture of contribution. See Case Study: Unlock the Value of Open Source (Fannie Mae).

Obstacles

- **Lack of developer buy-in —** Organizations that don't take developer pain points and strategic business goals into account while determining portal capabilities fail to see a return on their investment. A "build it and they will come" mindset can be a recipe for failure.

- **Ignoring the required integrations —** Portals that don't integrate with the software development life cycle tools for building, deploying and operating software struggle to increase adoption and stay relevant.

- **Absence of platform teams —** A dedicated platform team led by a product owner to manage and evolve the portal as a product is necessary to ensure the portal meets desired objectives. The absence of a dedicated product owner results in a disconnect between developer expectations and the portal's capabilities.

- **Treating Backstage as a ready-to-use portal —** Gartner regularly sees organizations misconstruing Backstage as a ready-to-use portal that requires no additional work. This sometimes leads to disillusionment, and the portal project is either put on the backburner or abandoned.

User Recommendations

- Pilot and implement portals by appointing a platform owner who can liaise with developers and build a roadmap to solve their most pressing challenges. Use metrics such as "platform coverage" to identify features that are needed and adopted widely.

- Assess open-source Backstage in comparison with commercial offerings by using criteria specific to your context: availability of skills, funding, urgency, hosting options, product capabilities across personas, support for integrations and most of all implementation effort.

- Provide consistent visibility and curated catalogs to help multiple teams and personas discover, access and use capabilities of underlying internal developer platforms.

Sample Vendors

Atlassian; configure8; Cortex; OpsLevel; OpsVerse; Port; Rely.io; Roadie; Solo.io; Spotify

Gartner Recommended Reading

Adopt Platform Engineering to Improve the Developer Experience

Market Guide for Internal Developer Portals

How to Implement Internal Developer Portals

Adopt Platform Engineering to Improve the Developer Experience

Top Strategic Technology Trends for 2024: Platform Engineering

AI-Augmented Software Engineering

Analysis By: Arun Batchu, Manjunath Bhat, Oleksandr Matvitskyy, Jim Scheibmeir

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Definition:

AI-augmented software engineering (AIASE) is the application of artificial intelligence technologies to assist software engineers throughout the software development life cycle. This includes the creation, validation, securing, deployment and maintenance of applications.

Why This Is Important

The software development life cycle involves routine tasks in both creative and operational DevOps loops. AI automation minimizes manual effort, freeing engineers to innovate, reduce technical debt, improve quality, security and team collaboration, and lower operational cost. AI technologies augment engineers' cognitive tasks such as analyzing logs, optimizing configurations, and generating scripts, code, unit tests and documentation.

Business Impact

Engineers using AI technologies, such as knowledge graphs, reinforcement learning, retrieval augmented generation and GPT-powered models trained on various data types, are more efficient and creative. AIASE accelerates application delivery by enabling software engineers to solve business problems and release solutions to production faster. AIASE helps engineers increase the nonfunctional quality properties of software such as security, performance and runtime costs.

Drivers

The drivers for demand include:

- The increasing complexity of software systems to be engineered.
- The increasing demand for developers to deliver high-quality code faster.
- The increasing number of security risks associated with software development.
- The need to optimize operational costs.

Drivers for providing the necessary technology solutions include:

- The application of AI models to prevent application security vulnerabilities by detecting them and suggesting fixes.
- The increasing impact of software development on business.
- The application of foundation models such as large language models (LLMs) to software code generation and optimization.
- The application of deep learning models to software operations.

Obstacles

- The hype surrounding this innovation has led to misconceptions and unrealistic expectations about the advantages of AIASE.
- Awareness about the tools ready for production is limited.
- Often, there's a lack of clarity about the origin and use of data for model training.
- Solutions are uneven and fragmented, automating only some tasks in the software development life cycle, such as code creation.
- There's a shortage of AI skills like prompt engineering, training, tuning, maintaining and troubleshooting models.
- High costs are associated with fine-tuning, training and inference of models at scale.
- Intellectual property risks arise from models trained on nonpermissive licensed code.
- Privacy issues arise from proprietary data and leaked code as training data for AI models.
- Technical employees fear that job automation by AI will lead to redundancy.

User Recommendations

- Pilot, measure and roll out tools only if there are clear gains.
- Innersource best practices, including examples of the best prompts to use for utilizing AIASE technology.
- 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 burdens.
- Reassure software engineers that AIASE is an augmentation toolset for human engineers, not a replacement.
- Choose providers (including open-source vendors) that provide visibility into 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

Amazon Web Services; Anima; CAST; Dynatrace; GitHub; GitLab; SeaLights; Sedai; ServiceNow; Veracode

Gartner Recommended Reading

How Platform Engineering Teams Can Augment DevOps With AI

Innovation Guide for AI Code Assistants

Top Strategic Technology Trends for 2024: AI-Augmented Development

Market Guide for AI-Augmented Software-Testing Tools

Predicts 2024: Generative AI Is Reshaping Software Engineering

Superapps

Analysis By: Tigran Egiazarov, Jason Wong

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

The superapp provides a unified mobile app experience to end users through a mobile platform with core features that fosters an ecosystem of miniapps. This ecosystem eradicates the requirement of a separate application marketplace and enables users to have a consistent and personalized app experience by giving the ability to discover, install and activate/deactivate miniapps inside the superapp.

Why This Is Important

Users demand mobile-first experiences that are powerful and easy to use. Consumer superapps have expanded beyond China and Southeast Asia to India (e.g., Tata Neu, MyJio and Paytm); Latin America (e.g., Rappi, PicPay, Mercado Libre); and the Middle East/Africa (e.g., M-PESA, Careem and Yassir). The superapp concept is rapidly expanding to employee-facing use cases, such as frontline workers; and employee communications and engagement, such as Walmart's me@Walmart and Wipro's MyWipro apps.

Business Impact

Organizations can create superapps to consolidate multiple mobile apps or related services that reduce user experience (UX) friction (such as context switching) and development effort. Superapps can achieve economies of scale and leverage the network effect of a larger user base and multiple providers. Superapps provide a more engaging experience for their customers, partners or employees. They improve UX by enabling users to activate their own toolboxes of miniapps and services.

Drivers

- Superapps are gaining interest from organizations that embrace composable application and architecture strategies to power new digital business opportunities in their industries or adjacent markets.

- Superapps are growing beyond mobile apps for consumer use cases. Frontline and remote work trends are driving the evolution of employee communications apps into workforce superapps through the addition of plug-ins for HR, payroll, shift management and other miniapp functions.

- The superapp concept is expanding into enterprise SaaS applications and tools, such as workflow, collaboration and messaging platforms (e.g., Slack and Microsoft Teams, which already have a large number of add-on apps to their main applications). Superapps are starting to expand to support a wide range of modalities, including chatbots, Internet of Things technologies and immersive experiences.

- A superapp advances the concept of a composite application by better integrating services, features and functions into a single app. Multiple internal development teams as well as external partners provide discrete services to users by building and deploying modular miniapps to the superapp, providing convenient access to a broader range of services in the app.

Obstacles

- Current technologies lack a ready-made framework or solution for creating custom superapps.

- Creating the business ecosystem can be a bigger challenge than technology implementation. A superapp serves as a platform for internally developed miniapps across the business and for third-party, externally developed miniapps. Business partners are needed to create an extended ecosystem for monetization by deploying miniapps to an established user base.

- Building the right UX for the right audience with design consistency between superapp and miniapp is hard. Different user personas prefer to interact differently with miniapps — for example, some may prefer single, task-focused miniapps versus others prefer everything at their fingertips. Inconsistent UX patterns in a superapp could negatively affect adoption and retention. Complications grow for workforce superapps when organizations convert multiple existing applications with varying UI and purposes under a single superapp by rebuilding them.

- Data sharing could be complex — including financial information, location data, and user preferences — and reinforced by simple user authentication, such as single sign-on (SSO), while tracking user preferences or app usage history. This raises concerns about data privacy and the potential for data breaches or personal information misuse. Ensuring security and governance is a crucial concern for the miniapp-superapp relationship.

- Obtaining publishing approval for the superapp from leading mobile device manufacturers is associated with several challenges, such as the legal battle between the Department of Justice and Apple over the restriction of superapps.

User Recommendations

- Identify whether a superapp is the right fit for your case using Gartner's decision tree framework (see Quick Answer: How Do We Decide If We Need a Superapp?).
- Educate partners on the innovations and business value a superapp strategy can drive to improve or consolidate mobile apps.
- Ensure there is a sound business model and organizational structure to support the distributed development ecosystem for miniapps.
- Secure executive sponsorship by preparing the partnership strategy and financial case for funding as a digital business initiative.
- Identify core features in your superapp (e.g., commerce, payments, communications or collaboration) that will drive a critical mass of adopters and create interest on the part of developers to serve those users.
- Choose a technology that supports real-time updates and extensibility without requiring a complete rebuild of the application, thereby minimizing time-to-market constraints.
- Offer an easy developer experience and convenient developer tools (e.g., APIs, design guidelines, software development kits and frameworks) for partners to build, test, register and submit miniapps into miniapp store for potential monetization.
- Define security and data protection needs by establishing governance reinforced with common platform implementation to satisfy security and data protection constraints.

Sample Vendors

Alipay; DingTalk; GeneXus; Ionic; KOBIL; Microsoft; PayPay; Paytm; Salesforce (Slack Technologies); WeChat

Gartner Recommended Reading

Quick Answer: What Is a Superapp?

Quick Answer: How Does a Superapp Benefit the Digital Employee Experience?

The Three Pillars of an Effective Workforce Superapp Strategy

Top Strategic Technology Trends for 2023: Superapps

Quick Answer: How Do We Decide If We Need a Superapp?

Cloud-Native

Analysis By: David Smith, Michael Warrilow

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

Cloud-native means something created to optimally leverage or implement cloud characteristics. Those characteristics, part of cloud computing's original definition, include being scalable and elastic, shared, metered by use, service-based, and ubiquitous by means of internet technologies. The term can be interpreted as taking full advantage of a provider's cloud capabilities, or using approaches pioneered in the cloud to deliver benefits wherever needed, via technologies such as containers.

Why This Is Important

"Cloud-native" is a popular term. Cloud-native is not one thing, and there is a battle of ideas that drives the multiple interpretations of the term.

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. 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 has similar issues. 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, yet may evolve into other technologies. This doesn’t guarantee portability, but is directionally consistent with the portability goal.

- Architectural needs can drive adoption. Examples include LIFESPAR and the Twelve-Factor App (i.e., cloud-native application design) and DevOps (cloud-native operations).

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. Be aware of any proprietary features you may use and of the consequences.

- 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.”

- 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.

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 cloud benefits.
- Assess vendor claims about their cloud-native capabilities with skepticism. Vendors use the term "cloud-native" to promote their offerings, regardless of what type of cloud-native meaning their offerings embody.
- Ensure 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

WebAssembly

Analysis By: Gregg Siegfried, Oleksandr Matvitskyy

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Definition:

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

Why This Is Important

Wasm can disrupt runtime environments like JavaScript, serverless or container runtimes. It improves software portability, efficiency, performance and security. As a World Wide Web Consortium (W3C) standard developed in partnership with vendors, Wasm now has web browser support. The server-side Wasm ecosystem is active, with standardization via the Cloud Native Computing Foundation (CNCF) and Bytecode Alliance. On X, a Docker co-founder said, “If WASM+WASI existed in 2008, we wouldn’t have needed to create Docker. That’s how important it is.”

Business Impact

Like the benefits from Java Virtual Machines, container runtimes and public cloud infrastructure, the benefits from Wasm are associated with technology. Thus, business impact is enabled by technology transformations like application replatforming and rearchitecting. Replacing containerized software with Wasm has the potential to profoundly reduce cloud compute resources (and costs). This approach supports sustainability objectives as well.

Drivers

- **Browser performance:** Modern browser-based UIs can be complex and heavy, with substantial application and presentation logic delegated to the client side. This complexity requires a runtime that is faster than interpreted JavaScript and able to deliver native or near-native compute performance.

- **Code portability:** The JavaScript Virtual Machine can be used on both the browser side and the server side. However, restricting browser-side implementation to JavaScript creates obstacles for developers proficient in other languages, and could hinder code sharing between the browser and server. Wasm allows development in the most popular languages for both the browser side and the 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 runtime allows extremely granular management of the “sandbox” where the code executes, thus minimizing 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, Wasm creates the request handlers when the requests are received, rather than keeping idle request handlers waiting for traffic.

- **Language flexibility:** Many programming languages can compile into Wasm. Rust is a particularly popular choice, but support is available today for JavaScript, Go, Python, C/C++ and others.

- **Kubernetes:** Platform software that simplifies the deployment and management of Wasm in Kubernetes, such as Cosmonic’s wasmCloud-operator and Fermyon’s SpinKube, is available and suitable for experimental use today.

Obstacles

- **Developer tooling:** Wasm represents a lower level of abstraction than the popular modern runtimes, so developers need improved tooling, component libraries and frameworks to keep development productivity high.
- **Architecture and skills:** While Wasm is designed for interoperability, architectural modifications are necessary before developers can compile their code directly for Wasm. Using Wasm in existing applications requires significant changes in application architecture and design.
- **Security risks:** Wasm supports much more granular control and can be safer than the JavaScript engine in browser implementations. However, in web browsers, the Document Object Model (DOM) currently relies on JavaScript, blocking opportunities to address security concerns through Wasm.
- **Toolchain maturity:** The DevOps toolchain for building, testing, deploying and releasing Wasm is progressing, but it is still best kept in the hands of experimenters and early adopters.

User Recommendations

- Pilot the use of Wasm for performance-sensitive client-side software, rather than defaulting to JavaScript. These pilots will acquaint product teams with the differences in developer experience and the language/toolchain requirements for incorporating Wasm into your stack.
- Prepare for the 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 organizations choosing to minimize the impact of Wasm on application architecture and implementation.
- Explore the server-side Wasm ecosystem, including the use of Wasm as a runtime for Kubernetes or at the content delivery network (CDN) edge. Encourage a small team to prototype with the platforms and tools available today.
- Track the evolution of Wasm as a toolchain for generative AI (GenAI) applications. Some developers consider it an ideal platform for this purpose, and interfaces like WASI-NN are advancing.

Sample Vendors

Cloud Native Computing Foundation; Cosmonic; Dylibso; Fastly; Fermyon; Google; Microsoft

Generative AI

Analysis By: Svetlana Sicular

Benefit Rating: Transformational

Market Penetration: 20% to 50% of target audience

Maturity: Adolescent

Definition:

Generative AI (GenAI) 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

GenAI exploration is widening:

- End-user organizations aggressively experiment with GenAI. Early adopters in most industries have had initial success with GenAI.
- Major technology vendors prioritize delivery of GenAI-enabled applications and tools.
- Numerous solutions have been emerging to innovate with foundation models, hardware and data for GenAI.
- Impacted by the GenAI hype, governments are introducing AI regulations and investing in national AI strategies.

Business Impact

Business focus is shifting from excitement around foundation models to use cases that drive ROI. Most GenAI implementations are currently low-risk and internal. With the rapid progress of productivity tools and AI governance practices, organizations will be deploying GenAI for more critical use cases in industry verticals and scientific discovery. In the longer term, GenAI-enabled conversational interfaces will facilitate technology commercialization, democratizing AI and other technologies.

Drivers

- Industry applications of GenAI are growing. GenAI reached creative work in entertainment, marketing, design, music, architecture and content generation.

- Best implementation practices are being discovered by the first enterprisewide deployments, and are fueling the top GenAI enterprise use cases: advanced chatbots, coding assistance and internal service desk. According to the 2023 Gartner AI in the Enterprise Survey, 18% of leaders highly involved in AI report that their organizations are advanced in GenAI adoption.

- GenAI is a top competitive area among major technology vendors. They compete on foundation model offerings, their enterprise readiness, pricing, infrastructure, safety and indemnification.

- New foundation models in new versions, sizes and capabilities are rapidly coming to market, making GenAI available for more use cases. Tools to improve model robustness, such as vector databases, graph technologies, LLM testing, security protection and open-source resources are making GenAI more usable.

- The progress is significant in multimodal models like Gemini or GPT4-Video, which are trained to take in both images and text; for example, they allow users to ask questions about images and receive answers via text. Models can combine concepts, attributes and styles to create original images, video and art, or translate audio to different voices and languages. Notably, text-to-image/video generation has advanced with the ability to create highly detailed and realistic visuals from textual descriptions.

- Enterprises are learning to use their own data with GenAI via prompt engineering and fine-tuning. AI-ready data and associated metadata have become central to GenAI strategies.

- Synthetic data helps enterprises to augment scarce data, mitigate bias, achieve superresolution or preserve data privacy.

- GenAI disrupts software engineering. Development automation techniques are promising to automate 5% to 10% of the programmers’ work. Organizations are now willing to tackle legacy modernization with GenAI.

Obstacles

- GenAI causes new ethical and societal concerns. Government regulations may hinder GenAI research. Pending regulations proliferate.
- Hallucinations, bias, a black-box nature and inexperience with a full AI life cycle preclude the use of GenAI in critical use cases for now.
- GenAI accountability, licensing and pricing are inconsistent among providers, and may catch customers off-guard.
- Reproducing results and finding references for generated information is challenging, but some validation solutions are emerging.
- Security professionals are new to certifying and protecting GenAI solutions; it will take time for security best practices to crystallize.
- GenAI is used for nefarious purposes. Full and accurate detection of generated content, such as deepfakes and disinformation, will remain challenging or impossible.
- The compute resources for training GenAI models are not affordable to most enterprises. Sustainability concerns about high energy consumption by GenAI are rising.

User Recommendations

- Identify low-risk use cases where you can improve your business with GenAI by relying on purchased capabilities. Consult vendor roadmaps to avoid developing similar solutions in-house.
- Architect your GenAI solutions to be ready for near-future upgrades, as foundation models and data tooling for them are progressing swiftly.
- Pilot ML-powered coding assistants, with an eye toward fast rollouts, to boost developer productivity.
- Use synthetic data to accelerate the development cycle and lessen regulatory concerns.
- Quantify the advantages and limitations of GenAI. Issue GenAI policies and guidelines, as it requires skills, funds and caution.
- Mitigate GenAI risks by working with legal, procurement, security and fraud experts. Technical, institutional and political interventions will be necessary to fight AI's adversarial impacts.
- Optimize the cost and efficiency of AI solutions by employing composite AI approaches to combine GenAI with other AI techniques.

Sample Vendors

Alibaba Cloud; Amazon Web Services; Anthropic; Google; Hugging Face; IBM; Meta; Microsoft; Mistral AI; OpenAI

Gartner Recommended Reading

Innovation Guide for Generative AI Technologies

Innovation Guide for Generative AI Models

How to Calculate Business Value and Cost for Generative AI Use Cases

When Not to Use Generative AI

10 Best Practices for Scaling Generative AI Across the Enterprise

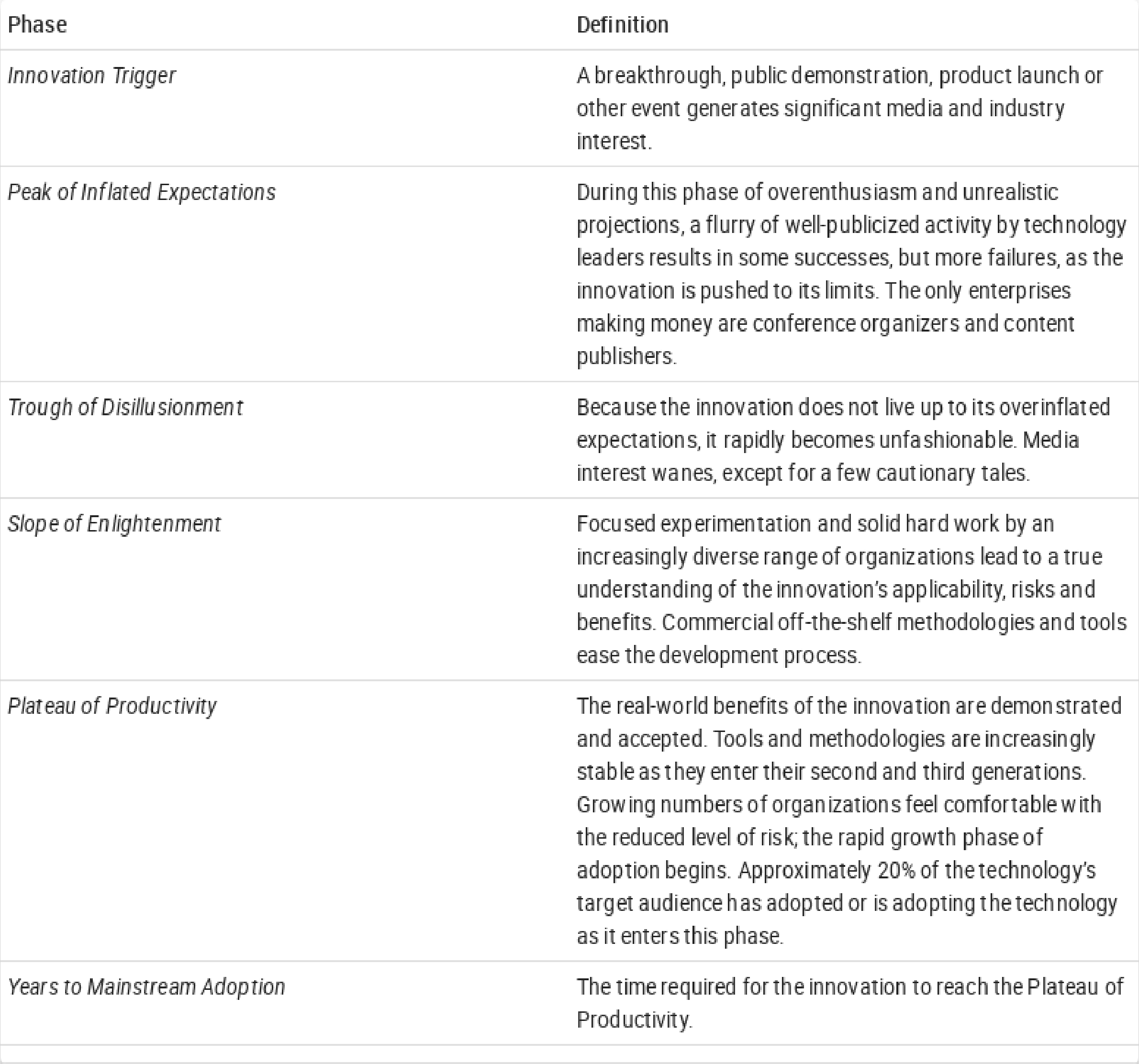
Appendixes

See the previous Hype Cycle: Hype Cycle for Emerging Technologies, 2023

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 2024)

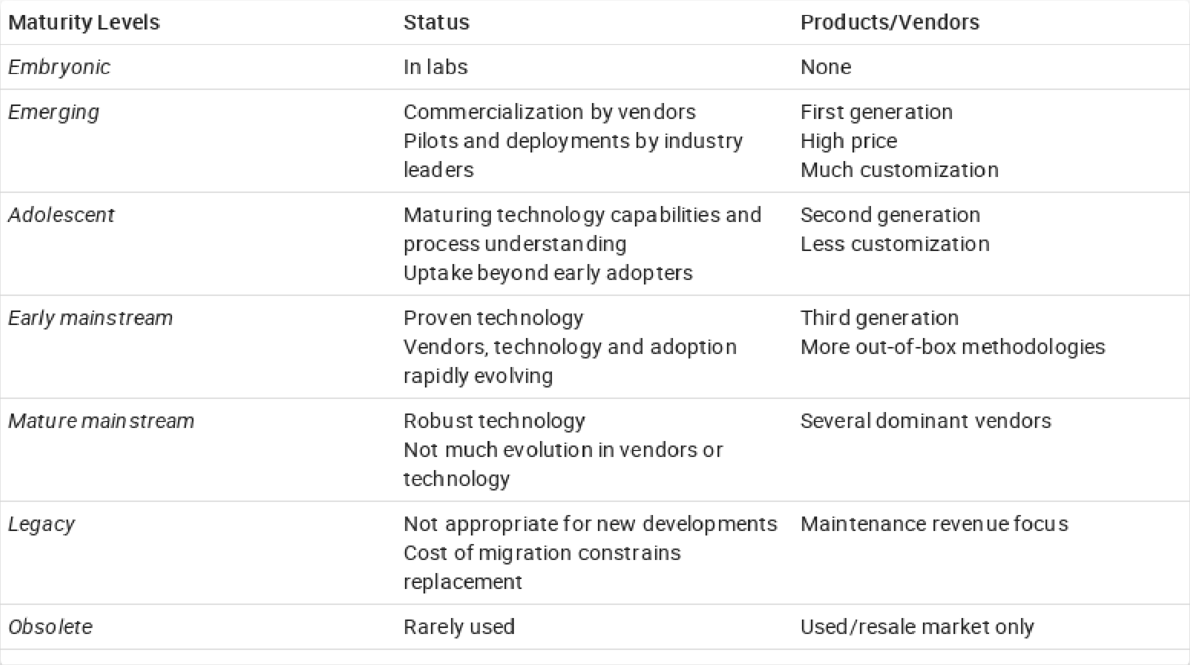
Table 3: Benefit Ratings

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 2024)

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 2024)

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 up till May 2024.

Methodology: Gartner conducts social listening analysis leveraging third-party data tools to complement or supplement the other fact bases presented in this document. Due to its qualitative and organic nature, the results should not be used separately from the rest of this research. No conclusions should be drawn from this data alone. Social media data in reference is from 1 January 2023 to 31 May 2024 in all geographies (except China) and recognized languages.

Document Revision History

Hype Cycle for Emerging Technologies, 2023 - 2 August 2023

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: Gartner's Hype Cycle Builder

Table 1: Priority Matrix for Emerging Technologies, 2024

Benefit	*Years to Mainstream Adoption*			
↓	***Less Than 2 Years*** ↓	***2 - 5 Years*** ↓	***5 - 10 Years*** ↓	***More Than 10 Years*** ↓
Transformational		AI-Augmented Software Engineering AI Supercomputing Generative AI WebAssembly	Autonomous Agents Digital Twin of a Customer Homomorphic Encryption	Artificial General Intelligence Cybersecurity Mesh Architecture Humanoid Working Robots Large Action Models
High	GitOps	AI TRiSM Disinformation Security Internal Developer Portals Prompt Engineering	Cloud-Native Digital Immune System Federated Machine Learning Multiagent Systems Reinforcement Learning Spatial Computing Superapps	6G Machine Customers
Moderate				
Low				

Source: Gartner (August 2024)

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.

Source: Gartner (August 2024)

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 2024)

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 2024)