Gartner Research

The Next Major Opportunity for Data and Analytics Services — Enabling Enterprise Generative AI

Ben Fieselmann, Radu Miclaus, Jim Hare

8 September 2023

Gartner®

The Next Major Opportunity for Data and Analytics Services — Enabling Enterprise Generative AI

Published 8 September 2023 - ID G00797396 - 13 min read

By Analyst(s): Ben Fieselmann, Radu Miclaus, Jim Hare

Initiatives: Product/Service Evolution and Management

As enterprises begin their generative AI journey, multiple points of friction will delay many high-quality outcomes. Offering managers should exploit services for data onboarding and enrichment, grounding application architectures and LLM data science support to accelerate GenAI adoption.

Overview

Key Findings

- Enterprises are asking how they can adopt generative AI (GenAI) in a secure manner, built on their private data, and within the context of use cases for internal users and customers.
- Enterprises have traditionally underinvested in knowledge management and information retrieval. Enterprises remain challenged in deploying traditional artificial intelligence applications, and GenAI adds new learning curves and strain on data science, support and operations teams.

Recommendations

Data and analytics offering managers looking to seize the disruptive enterprise opportunities caused by GenAI adoption initiatives should:

- Build offerings that show a path or roadmap to enterprisewide, secure GenAI adoption. Accelerate new content consumption experiences by including information retrieval, semantic indexing, and generative synthesis. Couple these with accelerators to efficiently and effectively augment the necessary and secure data enrichment and knowledge engineering activities.

- Develop offering options/features to drive clients to more rapid and effective adoption by specialists and general users. Provide ongoing/recurring, best approach training in applied usage and leverage to decrease learning curves, and the strain on data science, support and operations teams. Provide this training to also accelerate/increase adoption, and provide effective contextualization of data to achieve maximum ROI and dividends.
- Design and deliver specifically differentiated large language models (LLMs) and architecture-support reference architecture templates and assistance to reduce enterprise learning curves and hurdles in their internal/custom models and GenAI-powered applications.

Introduction

Mapping Service Offerings to Enterprise GenAI Readiness

Generative AI technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. Generative AI has profound impacts on business, including content discovery, creation, authenticity, and regulations, the automation of human work, and customer and employee experiences. The enterprise learning curve and low readiness for deploying trained productivity tools mean service providers can capitalize on their ability to sustain a faster learning curve than enterprises. Offerings managers will be challenged to formulate packaged accelerators (services and technology) offerings that gradually and appropriately map to the business outcomes that benefit from GenAI adoption for enterprises and their customers.

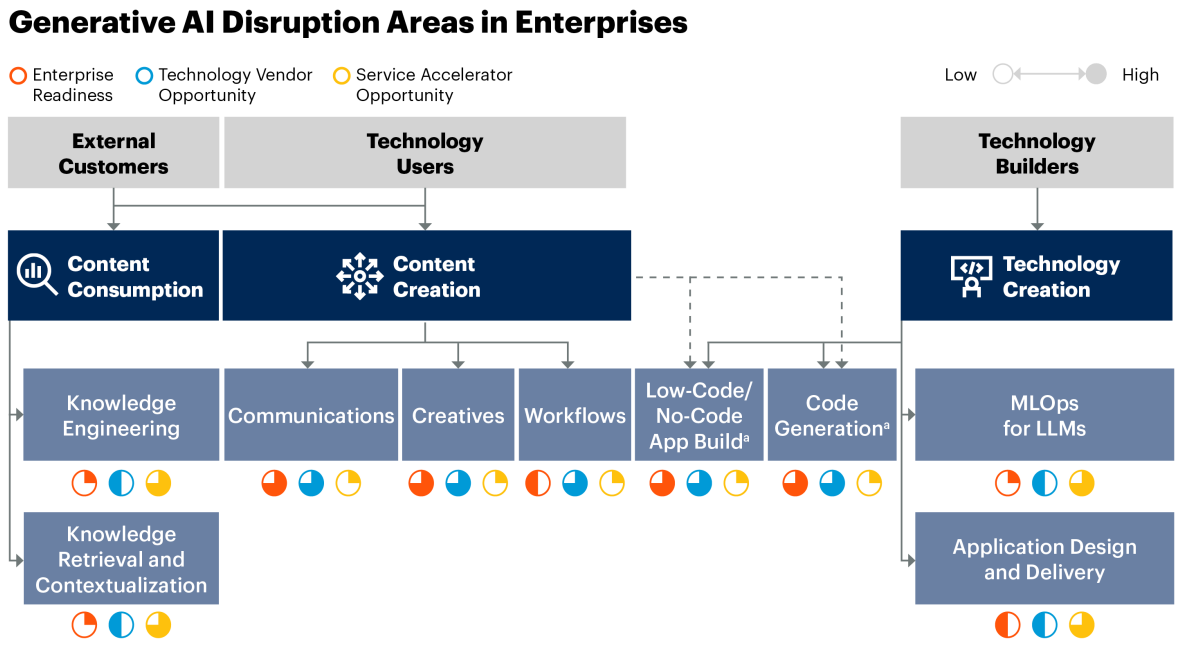
Generative AI is transforming the way humans interact with technology and disrupting technology usage patterns in the enterprise. In How TSPs Can Accelerate the Generative AI Readiness of Their Enterprise Customers, we describe content consumption, content creation and building applications as the main areas of disruption in the enterprise, and each one is met with a different readiness level by the enterprise. Enterprise skills gaps and need for agility are opening a wide area of opportunity for service providers, which we detail below.

- **Content consumption** in the context of enterprise knowledge manifested mostly in unstructured and semistructured data (policies, documentation, emails, contracts, codebases and other repositories) requires special considerations for acceptable usage. Generative content consumption needs to be grounded and informed by facts. This type of consumption assumes a new focus on knowledge retrieval based on role-based access control and building applications that use a hybrid architecture called retrieval augmented generation (RAG). [1] The skills gaps in the enterprise to meet the increasing needs for content consumption opens the door for service providers' targeted offerings.

- **Content creation** deals with the usage of productivity tools that are augmented by generative capabilities to speed up and scale output. The use cases encompass communications, creatives and design, business process and workflows, as well as low-code/no-code development. Service providers have an internal learning curve to develop differentiated skills in how to deliver outcomes using these tools faster and more efficiently than the enterprise in order to stay relevant.

- **Building applications** is another large opportunity for service providers (see Figure 1), as enterprises that want to build and/or fine-tune their own custom LLM and embedded generative applications will experience bottlenecks in skills, resources available and/or time to outcome. According to the 2021 Gartner AI in Organizations Survey (see Survey Analysis: The Most Successful AI Implementations Require Discipline, Not Ph.D.s), on average, only 54% of pilots make it to production, and it takes more than seven months to get to production. This opens the opportunity for service providers. Targeted offerings around support of model ops for LLMs, as well as design, delivery and scaling of generative AI applications, will enable service providers to add value fast and land and expand in enterprise accounts.

Figure 1. The Generative AI Disruption Areas in Enterprises and TSP Opportunities



Source: Gartner

[a] Low-code/no-code app build and code generation are content creation tools used by technology builders.

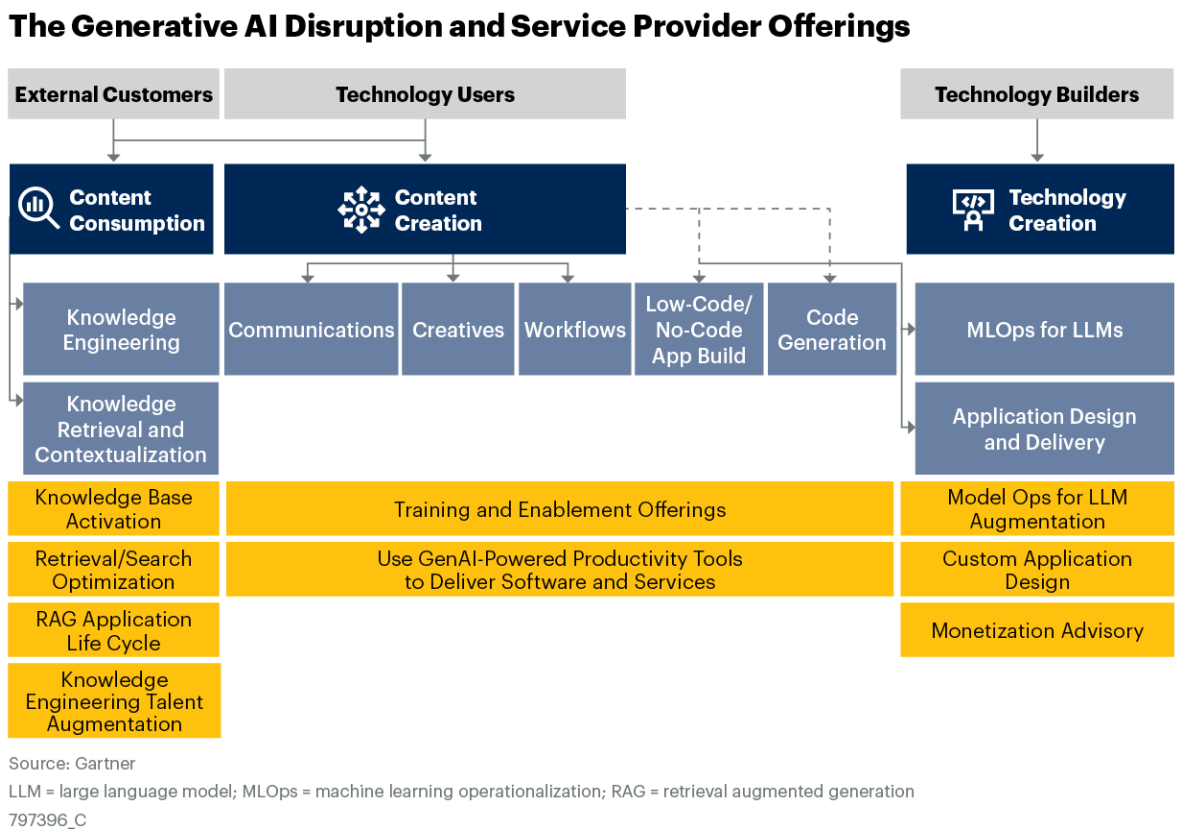
798693_C

Gartner

Analysis

In order to capitalize on the transformative opportunity of generative AI in the enterprise, service providers have to refine their current approach to AI service offerings and also sustain learning curves on how generative AI may disrupt their own usage patterns. Offering managers need to consider how offerings need to be formulated, factoring in enterprise readiness, as well as internal readiness for the delivery of outcomes in a predictable manner. (See Figure 2.)

Figure 2: The Generative AI Disruption and Service Provider Offerings



Source: Gartner
LLM = large language model; MLOps = machine learning operationalization; RAG = retrieval augmented generation
797396_C

Gartner

Glossary of Accelerators

Content Consumption Services Accelerators	*Content Creation Service Accelerators*	*Technology Creation Service Accelerators*
Knowledge Base Activation Accelerators	Training and Enablement Offerings	Model Ops for LLM Service Accelerators
Retrieval/Search Optimization		Custom Application Design, Build and Management
Retrieval Augmented Generation (RAG) Application Life Cycle		Monetization Advisory

Content Consumption Services Accelerators

As a summary from the readiness assessment for the enterprise on GenAI in How TSPs Can Accelerate the Generative AI Readiness of Their Enterprise Customers, most enterprises have a desire to use GenAI features on top of their private data and will expect the generative outputs to be grounded in internal factual information in order to be useful. This means that the generative process needs to be informed by internal information through a hybrid of search and generative synthesis. From a skills and process readiness perspective, the enterprise has traditionally underinvested in knowledge engineering skills (management, enrichment and retrieval of unstructured and semistructured data) and retrieval optimization skills. As a result, most enterprises will have a **low** readiness for building the RAG applications for content consumption and will benefit from services providers accelerators.

Knowledge Base Activation Accelerators

Enterprises maintain structured data in data stores. Procedural knowledge (policies, documentation, emails, contracts, etc.) is usually stored in business systems or document repositories (file systems, blob stores, etc.). Knowledge base activation entails a series of steps that get the knowledge bases aggregated and ready for exploration by knowledge workers.

Offerings for knowledge base activation need to include the following:

- Documentation of knowledge bases needed
- Aggregation of knowledge bases in a central location (optional)
- Indexing the knowledge bases and aggregation in central location
- Enrich/transform the knowledge base content with metadata tags, entity extraction, topics, embeddings/vectors building (for semantic search supported by vector databases), knowledge graph building (advanced)

Consultant skill sets needed: Data engineers with skill sets for knowledge engineering including tokenization, indexing, embeddings building, ontologies/taxonomies building, vector database knowledge, knowledge graph building.

Retrieval/Search Optimization

Once knowledge bases are activated, the search function needs to be built. Search complexity can vary from basic keyword search to more complex vector search (semantic search) or knowledge graph traverses.

Offerings for retrieval/search optimization need to include the following:

- Deployment of search engine in addition to the indexing done in previous steps
- Building the search pipelines (series of steps that interpret the incoming queries and match them with available documents)
- Integrate search pipeline steps with advanced AI models (including transformer-based models) and/or knowledge graphs
- Performance refinement for speed and scale for the profile of usage of the application

Consultant skill sets needed: Search engineering skills for search pipeline configuration, data science skills for retrieval models refinement and operations skills for performance optimization for the entire search pipeline.

Retrieval Augmented Generation (RAG) Application Life Cycle

RAG architectures are built by combining a search function with a generative service (like a chatbot) and works particularly efficiently for content consumption as well as communications content creation use cases. At a high level, the prompts are enhanced by retrieval results before being submitted to the generative AI chatbot, minimizing hallucinations and allowing for both generative synthesis and citation of factual sources. While technology vendors will offer the service to connect these services easily, service providers have a chance to layer additional assistance to refine these applications.

Offerings for RAG-based applications need to include the following:

- Security layer configuration at retrieval step based on the role-based access for the users. (Important step to confirm with enterprise information access controls).
- Vector database management/administration.
- Integration between search function and the generative chatbot application.
- Optimization of performance and cost structure of the application by iterating on scope and relevance of search results feeding into the prompting steps. This is done to minimize the size and efficiency of prompts to deliver accurate completions.
- Option for ongoing management as a service of the application including previous steps of continuous knowledge base activation and retrieval optimization.

Consultant skill sets needed: In addition to previous skill sets mentioned above, cloud engineering skills for services integration and ongoing monitoring would be needed.

Content Creation Service Accelerators

This area of disruption is about the productivity enhancement potential of generative AI assistants deployed directly in the productivity tools as copilots or assistants. The areas of substantial impact will be written communications, creative and design content (image, video voice, sound, etc.), as well as low-code/no-code generation for application development. The enterprise learning curve and **high** readiness for deploying trained productivity tools mean services providers can capitalize on their ability to sustain a faster learning curve than enterprises.

Training and Enablement Offerings

While the technology vendors will design the generative AI assistants experiences tailored to the context of the productivity tools, the learning curves will still be present for workers across the enterprise. Concepts related to best practices on prompt engineering and other aspects of operating these tools in a cost-effective manner can be opportunities for training and enablement content provided by service providers.

Offerings for training and enablement for interacting with AI-powered productivity tools should include:

- Initial skills gap assessments
- Learning plan development
- Training content delivery
- Train the trainer content and delivery

Consultant skill set needed: Tool- and task-specific skills based on the target area, curriculum planning development and delivery, communications and presentation skills.

Note: The learning curve for service providers with the new generative AI productivity tools needs to be sustained at a faster pace than the enterprise in order to stay ahead and offer value-added services in a cost-efficient manner. Service offering managers need to be aware of the consulting team's ability to deliver against specific use cases where productivity tools are essential in the output, in order to better formulate offerings targeting outcomes and scope accurately.

Technology Creation Service Accelerators

The enterprises that have a high level of AI ops maturity have access to large amounts of domain-specific data and/or which target the build of domain-specific applications using generative AI for content consumption or creation will sustain learning curves for developing custom LLMs and architecting applications. The LLMs skills talent pool is increasing, but still disproportionate to the massive demand. As a result, it is expected that enterprises will have skills gaps and **low** readiness both on the data science processes, as well as operations and application architecture domain. These areas represent large opportunities for service providers and well-defined service offerings as enterprises start assessing GenAI with new investment criteria (see Assess the Value and Cost of Generative AI With New Investment Criteria).

Model Ops for LLM Service Accelerators

The processes and methods for building and/or fine-tuning LLMs have some notable differences compared to traditional artificial intelligence/machine learning (AI/ML) processes. The notable differences are in the following areas:

- Data requirements for LLM training
- Infrastructure investment
- Development process, starting from base models and refining through prompt engineering and human feedback loops for GenAI model versus starting from scratch and hyperparameter tuning for traditional ML
- Different evaluation and monitoring techniques

These differences will create a skills gap and a learning curve for the enterprise. The service providers that develop expertise in fine-tuning models for different use cases and vertical domains, as well as managing the LLM ops end-to-end, will open themselves up to substantial opportunities in accelerating enterprises in custom generative AI application building.

Offerings for LLM ops services accelerators need to include:

- Infrastructure staging, including compute configurations and vector management facilities (vector libraries or databases)

- Best practices for base model selection-mapped to use cases and commercial considerations
- Data preparation for training and fine-tuning (Q&A pairs, synthetic data generation and data labeling techniques)
- Prompt engineering and management
- Evaluation and improvement interactions
- Best practices for using reinforcement learning with human feedback (RLHF)
- Documentations
- Deployment handoff

Consultant skill sets needed: LLM data science experience, linguistics experience, domain expertise for specialized models, operations and infrastructure staging expertise.

Custom Application Design, Build and Management

Understanding how generative AI back-end services should be integrated into applications, as well as optimization of other architecture components to support a scalable and cost-efficient application, is again a developing skills gap in the enterprise. Service providers that develop architecture design offerings to complement or augment the enterprise architecture teams during design time and even long-term management of applications will have an edge in meeting enterprise needs.

Offerings for generative AI applications design and management need to include:

- Mapping generative AI application needs to different patterns of deployment for LLMs (see AI Design Patterns for Large Language Models)
- Application architecture design and documentation- include architecture components like LLM stateless services, vector stores (APIs) for embeddings serving, compute cluster configuration, front-end API/webhooks
- Build planning and delivery (optional)
- Testing and QA

- Monitoring and observability
- Options for full delivery of GenAI-powered applications
- Options for (remote) management of GenAI-powered applications

Consultants skill sets needed: Architecture and solution engineering for AI applications, cloud services architects, full stack developers, automation engineering and operations.

Monetization Advisory

When building custom applications, enterprises will want to explore options for monetization of generative AI capabilities built using internal data and organizational knowledge. Beyond assistance to design and build the applications, service providers have the opportunity to accelerate the go-to-market motion for new enterprise offerings geared toward external monetization.

Offerings for monetization advisory should include:

- Product-market fit assessment
- Product packaging and pricing
- Messaging and launch planning and execution
- Sales enablement (as needed)
- Customer support function enablement

Consultants skill sets needed: Business and digital transformation, digital product life cycle management, AI product management, messaging, packaging and pricing, customer support function design and enablement.

Evidence

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, arXiv/Cornell University.

This research was informed by the following:

- Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI. This webinar was held on 30 March 2023 and 21 June 2023 with 1,465 and 1,079 respondents to the polling, respectively, for a total of 2,554 responses. Results of these polls should not be taken to represent all executives as the survey responses come from a population that had expressed interest in AI by attending a Gartner webinar on the subject.
- Gartner's The Future of ChatGPT and Generative AI in the Enterprise Webinar Poll, April 2023.
- 2021 Gartner AI in Organizations Survey — This survey was conducted to understand the keys to successful AI implementations and the barriers to the operationalization of AI. The research was conducted online from October through December 2021 among 699 respondents from organizations in the U.S., Germany and the U.K. Quotas were established for company size and for industries to ensure a good representation across the sample. Organizations were required to have developed AI or intended to deploy AI within the next three years. Respondents were required to be part of the organization's corporate leadership or report into corporate leadership roles, and have a high level of involvement with at least one AI initiative. Respondents were also required to have one of the following roles when related to AI in their organizations:

- Determine AI business objectives.
- Measure the value derived from AI initiatives.
- Manage AI initiatives development and implementation.

Disclaimer: Results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

Recommended by the Authors

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

Innovation Guide for Generative AI Technologies

Assess the Value and Cost of Generative AI With New Investment Criteria Emerging Tech: Top Use Cases for Generative AI

Assessing How Generative AI Can Improve Developer Experience

How to Pilot Generative AI

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

Content Consumption Services Accelerators	*Content Creation Service Accelerators*	*Technology Creation Service Accelerators*
Knowledge Base Activation Accelerators	Training and Enablement Offerings	Model Ops for LLM Service Accelerators
Retrieval/Search Optimization		Custom Application Design, Build and Management
Retrieval Augmented Generation (RAG) Application Life Cycle		Monetization Advisory

Actionable, objective insight

Position your organization for success. Explore these additional complimentary resources and tools for tech product management leaders:

eBook

3 Lessons From High-Growth Companies to Build a Successful Product Strategy

Find out how to avoid common product strategy pitfalls.

Download eBook

Report

Top Trends for Tech Providers for 2024

Explore the top trends product leaders must evaluate across all business dimensions.

Download Report

eBook

Leadership Vision for Technology Product Managers

Learn the top three strategic priorities for tech product managers.

Download eBook

Tool

Gartner Product Decisions

Inform your product strategy and roadmap with this tool.

Learn More

Already a client?
Get access to even more resources in your client portal. Log In

Connect With Us

Get actionable, objective insight that drives smarter decisions and stronger performance on your mission-critical priorities. Contact us to become a client:

U.S.: 1 844 466 7915

International: +44 (0) 3330 603 501

Become a Client

Learn more about Gartner for Product Teams
gartner.com/en/industries/high-tech

Stay connected to the latest insights

Attend a Gartner conference
View Conference

Gartner