Generative AI workload assessment

AWS Prescriptive Guidance



Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Prescriptive Guidance: Generative AI workload assessment

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Services or capabilities described in AWS documentation might vary by Region. To see the differences applicable to the AWS European Sovereign Cloud Region, see the AWS European Sovereign Cloud User Guide.

Table of Contents

Introduction **1**
Purpose of this guide 1
Target audience and benefits 2
Scope 2
Targeted business outcomes **4**
Assessment considerations and prerequisites **7**
Start with clear use cases 7
Ensure business alignment 8
Implement governance and oversight 8
Address data and technical prerequisites 8
Consider compute resource requirements 8
Address privacy and security implications 9
Engage stakeholders early 9
Iterate and learn 9
Generative AI workload assessment questionnaire **10**
Readiness 10
Use cases 12
Architecture 16
Storage 16
Regulations and compliance 18
Integration 18
Testing 20
Deployment and automation 22
Data strategy 24
Translating assessment insights into actionable outcomes **27**
Next steps **29**
FAQ **30**
What is the primary objective? 30
Who should use this assessment? 30
What are the key components? 30
How does this help define the architecture? 30
What are the benefits? 31
How can we implement this successfully? 31
What are the challenges? 31

What are the regulatory and compliance requirements? 31
What is the role of stakeholders? 31
How can we measure success? 32
How does the approach differ based on organization size? 32
Resources 33
Document history 34

Generative AI workload assessment

Tabby Ward and Deepak Dixit, Amazon Web Services (AWS)

November 2024 (document history)

Generative AI workload assessment is a strategic method aimed at evaluating and improving an organization's preparedness to create or update its generative AI workloads. This assessment is important because incorporating generative AI into business operations can greatly change how things work, and can provide new efficiencies and capabilities. However, to adopt generative AI successfully, it's essential to thoroughly understand current systems and have a clear plan for the future.

Generative AI workloads refer to computational tasks that involve the use of artificial intelligence models that can create new content, such as text, images, code, or other data types. These workloads typically require substantial computing power, specialized hardware such as GPUs, and large datasets for training and inference. Integrating generative AI workloads into operations presents several challenges:

- Infrastructure requirements: Provisioning the significant computational resources and specialized hardware that generative AI models require.
- Data management: Ensuring data quality, privacy, and compliance while handling large datasets.
- Skills gap: Lack of expertise in AI technologies and model deployment.
- Ethical considerations: Addressing bias, fairness, and transparency in AI-generated content.
- Integration complexity: Seamlessly incorporating generative AI into existing workflows and legacy systems.
- Cost management: Balancing the potential benefits with the high costs of implementation and operation.

Overcoming these challenges requires careful planning, investment in infrastructure and talent, and a strategic approach to implementation.

Purpose of this guide

Generative AI is rapidly becoming a critical component across many industries. It provides transformative opportunities but also pose challenges in terms of integration, compliance, and

scalability. Many organizations struggle to fully leverage AI due to weak technological foundations, resistance to change, and data quality issues. The generative AI workload assessment addresses these challenges by identifying the requirements for modernization, defining the scope of implementation, and challenging legacy systems and thinking. It also aids in determining minimum viable products (MVPs) and helps you develop a target solution architecture, ensuring a structured and strategic approach to AI adoption.

This guide serves as a structured approach to help organizations navigate the complexities of adopting generative AI technologies. Instead of clearly defining requirements from the outset, the guide assists in:

- Identifying potential use cases for generative AI within your organization.
- Assessing your organization's readiness for generative AI adoption.
- Defining and refining use case goals and stretch goals.
- Determining the scope and requirements for generative AI implementation.
- Developing a target solution architecture.

Target audience and benefits

This assessment is specifically designed for solutions architects, enterprise architects, and application architects who want to evaluate the technical aspects of generative AI workload modernization. It is also valuable for program and people managers who want to gauge their team's overall readiness, resource allocation, and enablement requirements. Industry best practices emphasize the importance of a comprehensive assessment to ensure readiness for AI adoption. This includes evaluating architecture, storage, compliance, integration, testing, deployment, and automation.

Scope

The following topics are in-scope for the generative AI workload assessment method:

- Current generative AI technologies and models (for example, large language models, image generation models)
- Narrow AI applications that use generative techniques
- Integration of generative AI with existing systems and workflows
- Data strategies for training and fine-tuning generative AI models

- Ethical considerations and responsible AI practices for current generative AI applications
- Testing and deployment strategies for generative AI in production environments
- Security and privacy considerations for generative AI implementations
- Performance optimization and scalability of generative AI workloads
- Use cases and applications of generative AI in various industries
- Evaluation of generative AI outputs and quality assurance processes

The following topics are out of scope:

- Artificial general intelligence (AGI) and artificial superintelligence (ASI) scenarios
- Speculative future advancements in AI beyond current generative models
- Quantum computing applications in AI
- Neuromorphic computing and brain-computer interfaces
- Consciousness and self-awareness in AI systems
- Long-term societal impacts of advanced AI beyond current generative AI applications
- Regulatory frameworks for hypothetical future AI technologies
- Philosophical debates on the nature of intelligence and consciousness in machines
- Extreme edge cases or highly speculative use cases of AI
- Detailed technical specifications of proprietary AI models or architectures

Targeted business outcomes

The generative AI workload assessment aims to deliver several targeted outcomes that are crucial for successfully modernizing generative AI workloads. These outcomes ensure that organizations are well prepared to integrate AI technologies effectively and efficiently.

For each targeted outcome, the generative AI workload assessment focuses on:

- Inter-dependencies: Identify and clarify any inter-dependencies between the outcome and other aspects of the modernization process. This includes understanding how one outcome might influence or be influenced by others, to ensure a holistic approach to modernization.
- Stakeholder alignment: Outline strategies to align various stakeholders with each outcome. This involves communicating the value and impact of each outcome to different organizational levels and departments, to foster buy-in and support.
- Prioritization: In cases where multiple use cases or outcomes are identified, provide a framework for prioritizing them based on factors such as business impact, resource requirements, and strategic alignment.
- Continuous improvement: For each outcome, establish mechanisms for ongoing evaluation and refinement. This ensures that the modernization efforts remain adaptive and responsive to changing technological landscapes and business needs.

Here is a detailed discussion of each targeted outcome:

Target architecture

- Definition: The assessment helps define a clear and scalable target architecture for generative AI workloads.
- Components: This includes selecting appropriate cloud services, designing data pipelines, and ensuring system interoperability.
- Benefits: A well-defined architecture supports scalability, reliability, and performance optimization, and provides a strong foundation for modernization.

Customer readiness

- Evaluation: Assess the current state of the organization's infrastructure, processes, and culture to determine readiness for generative AI modernization adoption.

- Criteria: This involves evaluating technical capabilities, data quality, and organizational willingness to embrace change.
- Outcome: Identifying gaps and areas for improvement ensures that the organization is prepared for a smooth transition to modern solution and technologies.

Use case goals and stretch goals

- Use case goals establish clear objectives for target solution implementation, focusing on specific business problems or opportunities.

  A use case goal in the context of generative AI modernization refers to a specific, measurable objective that an organization aims to achieve by implementing generative AI solutions. These goals are typically aligned with broader business objectives and focus on addressing particular challenges or opportunities within the organization. Examples of use case goals might include:

  - Reducing customer service response time by 50 percent by using generative AI-powered chatbots.
  - Improving code review efficiency by 30 percent through generative AI-assisted code analysis.
  - Enhancing fraud detection accuracy by 25 percent by using generative AI pattern recognition.
- Stretch goals define ambitious targets that push the boundaries of what generative AI modernization can achieve within the organization.
- Impact: Setting both achievable and aspirational goals helps align generative AI modernization initiatives with strategic business objectives and encourages innovation.

Effort estimation

- Purpose: Accurate effort estimation aids in resource planning and ensures that projects are delivered on time and within budget.
- Scope: Estimate the resources, time, and budget required to implement the generative AI modernization plan.
- Factors: Consider technical complexity, integration challenges, and potential risks.

Enablement needs

- Training and development: Identify the skills and knowledge required for successful generative AI modernization adoption.

- Resources: Determine the need for training programs, workshops, and other enablement activities.
- Outcome: Ensuring that staff are equipped with the necessary skills enhances the effectiveness of generative AI modernization initiatives and supports long-term success.

Implementation plan

- Roadmap: Develop a detailed plan that outlines the steps required to achieve generative AI modernization.
- Milestones: Define key milestones and deliverables to track progress.
- Benefits: A clear implementation plan provides direction and accountability, and facilitates a structured approach to generative AI modernization.

Assessment considerations and prerequisites

Start with clear use cases

Identify specific business problems or opportunities that generative AI can address. Focus on use cases that align with strategic business goals and offer measurable benefits. Prioritize use cases that target commonly faced challenges within the organization to ensure that the solution architecture can serve as a pattern for multiple scenarios.

Initiating the assessment process with a general understanding of potential generative AI applications is beneficial but not mandatory. The questionnaire that's included with this guide accommodates various levels of preparedness, from organizations that have well-defined use cases to those that have only broad ideas. The assessment process serves to:

- Refine and clarify these initial use case ideas.
- Identify new potential use cases.
- Develop specific, measurable goals for each use case.
- Assess the feasibility and potential impact of each use case.

Let's consider a hypothetical example: A financial services company decides to explore generative AI modernization. They start with a broad idea of improving their customer service and fraud detection processes.

- **Initial assessment**: The questionnaire helps them evaluate their current systems, data quality, and organizational readiness for generative AI adoption.
- **Use case refinement**: Through the assessment process, they refine their initial ideas into two specific use cases:
  - Implementing a generative AI-powered chatbot for customer inquiries
  - Using generative AI for real-time transaction fraud detection
- **Goal setting**: For each use case, they define specific goals:
  - Reduce customer service response time by 40 percent within 6 months
  - Improve fraud detection accuracy by 20 percent and reduce false positives by 15 percent
- **Stretch goals**: They also set these ambitious targets:

  - Achieve 80 percent customer satisfaction with AI-assisted responses
  - Develop a predictive fraud detection model that identifies new fraud patterns
- **MVP definition**: The questionnaire helps them determine an MVP for each use case, focusing on essential features that deliver immediate value.
- **Target architecture**: Finally, they develop a target architecture that supports one or both use cases, and ensures scalability and integration with existing systems.

Ensure business alignment

Align generative AI initiatives with overall business strategy and objectives. For each use case, develop a clear value proposition that demonstrates how generative AI contributes to business growth, efficiency, or innovation. Establish metrics to measure the impact of generative AI implementations on key performance indicators (KPIs).

Implement governance and oversight

Create a cross-functional steering committee to oversee generative AI initiatives. Develop policies and guidelines for responsible AI use, addressing ethical considerations and potential biases. Establish a review process for generative AI projects to ensure compliance with organizational standards and regulatory requirements.

Address data and technical prerequisites

Assess and improve data quality, and implement data governance practices to ensure reliable inputs for generative AI models. Develop a data strategy that addresses data collection, storage, and management that are specific to generative AI needs. Evaluate and enhance data infrastructure to support the volume and velocity of data required for generative AI workloads.

Consider compute resource requirements

Assess current IT infrastructure and identify gaps in computational capacity for generative AI workloads. Plan for scalable compute resources, considering options such as cloud services or on-premises high-performance computing clusters. Optimize resource allocation to balance performance and cost-effectiveness for both training and inference workloads.

Address privacy and security implications

Implement robust security measures to protect sensitive data used in generative AI training and operations. Ensure compliance with data protection regulations such as General Data Protection Regulation (GDPR) or California Consumer Privacy Act (CCPA) when handling personal information. Develop protocols for secure model deployment and monitoring to prevent unauthorized access or misuse of generative AI capabilities.

Engage stakeholders early

Involve key stakeholders from the beginning to gain leadership buy-in and support. Clearly communicate the benefits and potential impact of modernization initiatives, specifically for generative AI workloads. Provide training and resources to help stakeholders understand generative AI technologies and their implications.

Iterate and learn

Adopt an incremental approach that lets you refine target solutions. Use feedback loops to continuously improve workload architecture and processes. Regularly assess the performance and impact of generative AI implementations, and adjust strategies as needed based on real-world results and evolving business needs.

Generative AI workload assessment questionnaire

The following sections provide questions that you can use to evaluate different aspects of generative AI workload modernization for your organization. This comprehensive questionnaire evaluates your organization's readiness to adopt and implement generative AI workloads with questions across key areas, including use cases, architecture, storage, compliance, integration, testing, deployment, and data strategy. By addressing critical aspects of generative AI implementation, from technical infrastructure to regulatory considerations, this questionnaire helps you identify strengths, gaps, and opportunities in your AI modernization journey.

Sections:

- Readiness
- Use cases
- Architecture
- Storage
- Regulations and compliance
- Integration
- Testing
- Deployment and automation
- Data strategy

You can also download the questionnaire in Microsoft Excel format and use it to record your information.

Download questionnaire

Readiness

Question	Example response
Do you have AWS accounts that can be leveraged for these workloads?	Yes or no.

Question	Example response
Do you have an existing enterprise agreement with AWS?	Yes or no.
How scalable is your current cloud infrastru cture to handle generative AI workloads?	Our cloud infrastructure is highly scalable, with automatic scaling capabilities for compute resources and distributed storage systems that are designed to handle large-sca le generative AI workloads efficiently.
Do you have data pipeline capabilities for preprocessing and feature engineering at scale?	Our data pipelines use distributed processing frameworks such as Apache Spark for large-scale data preprocessing and feature engineeri ng, with support for both batch and streaming data processing.
Do you have account provisioning and management capability?	Yes or no.
How would you describe your organization's AI literacy and readiness to adopt generative AI technologies?	Our organization has invested heavily in AI education programs, and most technical staff has completed basic AI/ML training. The organization has a culture of innovatio n that embraces new technologies, including generative AI.
What AI/ML expertise exists within your organization, and how is it distributed?	We have a dedicated AI Center of Excellenc e with experienced data scientists and ML engineers. We upskill domain experts across different business units to become AI-literate and to identify generative AI use cases.
Do you have a high-level business case that articulates the cloud program objectives, benefits, and cost?	Yes or no.

Question	Example response
What is your time line to take the solution to production?	Weeks, months, and so on.
Has a funding commitment been made by your key stakeholders (for example, CFO, CIT/CTO, COO)?	Yes or no.
How do you ensure compliance with data protection regulations in your generative AI initiatives?	We have a dedicated compliance team that works closely with our AI teams. We conduct regular privacy impact assessments, implement data protection by design principles, and maintain detailed data processing records for all generative AI projects.
How mature are your existing systems that integrate with new generative AI technologies?	Our IT architecture is based on microservices and APIs that allow for flexible integration of new generative AI technologies. These systems are standardized on common data formats and protocols to ensure interoperability.
What experience do you have in operationalizing ML models, and how might this apply to generative AI systems?	We have established MLOps practices, including automated model deployment pipelines, monitoring systems, and A/B testing frameworks. These practices are being adapted to handle the unique requirements of large-scale generative AI models.

Use cases

Question	Example response
What is the primary goal or success criteria of the use case?	To improve customer support response time, increase sales conversions, enhance product recommendations. Also: To improve user

Question	Example response
	satisfaction, task completion rate, response quality, and so on.
How does this use case align with your organization's strategic goals?	This aligns with our strategic goal of enhancing customer satisfaction by reducing response times in customer service.
What is the expected volume of data or requests for the use case?	500 transactions per second (TPS).
What types of data sources are required to support your generative AI workloads?	Internal structured databases (customer records, sales data, and so on); unstructured text data from documents, emails, and social media; audio and video files for speech and image recognition tasks; real-time streaming data from IoT devices and sensors; public datasets and APIs for enrichment.
How frequently do you need to update or refresh data from these sources?	Transactional databases: near real-time updates; document repositories: daily batch updates; social media feeds: hourly updates; IoT sensor data: continuous real-time streaming; public datasets: monthly or quarterly updates.
What data formats do your generative AI models require as input?	Structured data: CSV, JSON, and SQL database tables; text data: plain text, PDF, and HTML; image data: JPEG, PNG, and TIFF; audio data: WAV and MP3; video data: MP4 and AVI.

Question	Example response
What are your key data quality concerns for generative AI workloads?	Completeness: ensuring that no critical fields are missing; accuracy: verifying data correctness and eliminating errors; consisten cy: maintaining uniform formats and values across sources; timeliness: ensuring that data is up to date for real-time inference; relevance : confirming that data aligns with the specific generative AI task.
What are the key performance requireme nts (for example, response time, throughput, accuracy)?	95% accuracy; < 500 ms response time; ability to handle 1000 requests/sec. High accuracy (95%+), moderate accuracy (80-90%), best effort, and so on.
Do you have any other KPIs to measure the success of this use case ?	Key KPIs include error rate reduction, time savings per transaction, and customer satisfact ion scores.
How much model accuracy is desired, and how does it balance with the cost?	High accuracy (>90%) with moderate cost, moderate accuracy (70-80%) with low cost, and so on.
What are the primary use cases or scenarios for the generative AI solution?	Customer service chatbot, content generation, product recommendation, and so on.
What are the target users or personas for the generative AI system?	Customer service agents, marketing team,employees, end users, and so on.
What is the expected volume of requests or users?	1,000 requests per day; 10,000 monthly active users.
Are there any specific use case constraints or requirements?	Real-time response, multi-lingual support, data privacy, and so on.
Do you have an allocated budget for developin g and maintaining the generative AI solution?	The initial development cost is estimated at $200,000, with annual maintenance costs of $50,000.

Question	Example response
What is the projected return on investment (ROI) and payback period for this use case?	Expected ROI of 150% over three years, with a payback period of 18 months.
Are there any hidden costs or potential savings that should be considered?	Potential savings include reduced overtime costs. Hidden costs might involve additional training for staff.
What are the scalability and future expansion possibilities of this generative AI solution?	The solution is designed to scale with our operations, with the possibility of expanding to other departments in the future.
How do you ensure fairness and mitigate bias in your generative AI models?	We plan to mitigate bias through diverse data collection, regular bias audits, and implement ation of bias mitigation techniques.
What processes do you have in place for addressing ethical concerns or unintended consequences?	We will manage ethical concerns through an established AI incident response plan, regular ethical risk assessments, an anonymous reporting system for employees, collaboration with external ethics experts, and continuou s monitoring and adjustment of deployed models based on feedback.
How do you approach prioritizing and sequencing generative AI workload assessmen ts across different projects and departments in your organization?	By conducting a high-level survey across all departments to identify potential generativ e AI use cases and evaluating them based on three key criteria: business impact, technical feasibility, and ethical considerations. Projects with high potential impact, lower technical barriers, and minimal ethical concerns are given priority.

Architecture

Question	Example response
What type of generative AI model or architecture is being considered?	Transformer, convolutional neural network (CNN), recurrent neural network (RNN), decision trees, and so on.
What is the expected scale or volume of data and computations?	Millions of users, petabytes of data, and so on.
What are the hardware requirements (for example, CPUs or GPUs) for training and inference?	High-end GPUs, CPU clusters, cloud instances, and so on.
How will the generative AI model be updated or retrained over time?	Through continuous learning, periodic retraining, manual updates, and so on.
What are the data preprocessing and feature engineering requirements?	Text cleaning, image augmentation, feature selection, and so on.
How will the generative AI system handle edge cases, outliers, or low-confidence inputs?	Through fallback to human oversight, request clarification, and so on.
What are the latency requirements for the generative AI application?	Real-time, near real time, batch processing, and so on.

Storage

Question	Example response
Where will the training data be stored?	In cloud storage (for example, Amazon S3, file storage, block storage, or object storage), in on-premises storage, and so on.

Question	Example response
What are the storage requirements for the training data and model artifacts (for example, capacity, durability, availability)?	Petabyte-scale storage, high durability (99.999999999% durability), high availability, and so on.
What are the data retention and backup requirements for the training data and model artifacts?	Data retention for *x* years, daily backups, off-site backups, and so on.
Which file formats are primarily used for storing your AI training datasets (for example, CSV, JSON, Parquet, HDF5)?	Parquet files for structured data, and HDF5 for large multidimensional arrays and unstructu red data such as images and text. We use specialized formats such as TFRecord to optimize data loading during training.
How are your training datasets organized: as individual files, in databases, or using specializ ed AI data formats?	Small to medium datasets are stored as individual Parquet files in object storage for flexibility. Large datasets are stored in a distributed database (Cassandra) to handle scale.
Do you use any data compression or encoding techniques specifically for generative AI training data?	For tabular data, we use dictionary encoding and bit-packing techniques that are available in Parquet. For images, we use lossy JPEG compression with quality settings optimized for our models.
How do you handle versioning and storage of different iterations of training datasets? What impact does this have on your overall storage needs?	We use a data versioning system (DVC) that is integrated with our ML platform.

Regulations and compliance

Question	Example response
What are the relevant regulations or compliance requirements for the generative AI solution (for example, GDPR, HIPAA, PCI-DSS)?	GDPR for handling personal data, HIPAA for healthcare data, PCI-DSS for payment data, and so on.
What ethical generative AI guidelines or frameworks has your organization adopted?	We implemented our own responsible AI guidelines. All generative AI projects undergo ethical review before approval and deploymen t.
What are the security requirements for the generative AI system?	Data encryption, secure network communica tion, regular security audits.
What are the requirements for data privacy and protection?	Data anonymization, encryption, access control, and so on.
What are the requirements for the solution to handle sensitive or confidential data?	Strict access controls, data masking, data residency requirements, and so on.
How will user authentication and authoriza tion be handled?	By using OAuth, API keys, single sign-on (SSO), and role-based access control (RBAC).
How will the solution be monitored and managed in production?	By using monitoring tools such as Prometheus and Datadog, logging tools such as ELK Stack, alerting systems, and so on.

Integration

Question	Example response
What are the requirements for integrating the generative AI solution with existing systems or data sources?	REST APIs, message queues, database connectors, and so on.

Question	Example response
How will data be ingested and preprocessed for the generative AI solution?	By using batch processing, streaming data, data transformations, and feature engineeri ng.
How will the output of the generative AI solution be consumed or integrated with downstream systems?	Through API endpoints, message queues, database updates, and so on.
Which event-driven integration patterns can be used for the generative AI solution?	Message queues (such as Amazon SQS , Apache Kafka, RabbitMQ), pub/sub systems, webhooks, event streaming platforms.
Which API-based integration approaches can be used to connect the generative AI solution with other systems?	RESTful APIs, GraphQL APIs, SOAP APIs (for legacy systems).
Which microservices architecture component s can be used for the generative AI solution integration?	Service mesh for inter-service communication, API gateways, container orchestration (for example, Kubernetes).
How can hybrid integration be implemented for the generative AI solution?	By combining event-driven patterns for real-time updates, batch processing for historical data, and APIs for external system integration.
How can the generative AI solution output be integrated with downstream systems?	Through API endpoints, message queues, database updates, webhooks, and file exports.
Which security measures should be considered for integrating the generative AI solution?	Authentication mechanisms (such as OAuth or JWT), encryption (in transit and at rest), API rate limiting, and access control lists (ACLs).
How do you plan to integrate open source frameworks such as LlamaIndex or LangChain into your existing data pipeline and generative AI workflow?	We're planning to use LangChain to build complex generative AI applications, particula rly for its agent and memory managemen t capabilities. We aim to have 60% of our generative AI projects using LangChain within the next 6 months.

Question	Example response
How will you ensure compatibility between your chosen open source frameworks and your existing data infrastructure?	We're creating a dedicated integration team to ensure smooth compatibility. By the third quarter, our goal is to have a fully integrate d pipeline that uses LlamaIndex for efficient data indexing and retrieval within our current data lake structure.
How do you plan to leverage the modular components of frameworks such as LangChain for rapid prototyping and experimentation?	We're setting up a sandbox environment where developers can quickly prototype by using LangChain's components.
What is your strategy for keeping up with updates and new features in these rapidly evolving open source frameworks?	We've assigned a team to monitor GitHub repositories and community forums for LangChain and LlamaIndex. We plan to evaluate and integrate major updates quarterly, with a focus on performance improvements and new capabilities.

Testing

Question	Example response
What are the testing requirements (for example, unit testing, integration testing, end-to-end testing)?	Unit testing for individual components, integration testing with external systems, end-to-end testing for critical scenarios, and so on.
How do you ensure data quality and consisten cy across different sources for generative AI training?	We maintain data quality through automated data profiling tools, regular data audits, and a centralized data catalog. We've implemented data governance policies to ensure consistency across sources and to maintain data lineage.
How will the generative AI model be evaluated and validated?	By using a holdout dataset, human evaluation, A/B testing, and so on.

Question	Example response
What are the criteria for evaluating the performance and accuracy of the generative AI model?	Precision, recall, F1 score, perplexity, human evaluation, and so on.
How will edge cases and corner cases be identified and handled?	By using a comprehensive test suite, human evaluation, adversarial testing, and so on.
How will you test for potential biases in the generative AI model?	By using demographic parity analysis, equal opportunity testing, adversarial de-biasing techniques, counterfactual testing, and so on.
Which metrics will be used to measure fairness in the model's outputs?	Disparate impact ratio, equalized odds, demographic parity, individual fairness metrics, and so on.
How will you ensure diverse representation in your test datasets for bias detection?	By using stratified sampling across demograph ic groups, collaboration with diversity experts, use of synthetic data to fill gaps, and so on.
Which process will be implemented for ongoing monitoring of model fairness post-deployment?	Regular fairness audits, automated bias detection systems, user feedback analysis, periodic retraining with updated datasets, and so on.
How will you address intersectional biases in the generative AI model?	By using intersectional fairness analysis, subgroup testing, collaboration with domain experts on intersectionality, and so on.
How will you test the model's performan ce across different languages and cultural contexts?	By using multilingual test sets, collabora tion with cultural experts, localized fairness metrics, cross-cultural comparison studies, and so on.

Deployment and automation

Question	Example response
What are the requirements for scaling and load balancing?	Intelligent request routing; automatic scaling system; optimizing for fast cold starts by employing techniques such as model caching, lazy loading, and distributed storage systems; designing the system to handle bursty, unpredictable traffic patterns.
What are the requirements for updating and rolling out new versions?	Blue/green deployments, canary releases, rolling updates, and so on.
What are the requirements for disaster recovery and business continuity?	Backup and restore procedures, failover mechanisms, high availability configurations, and so on.
What are the requirements for automating the training, deployment, and management of the generative AI model?	Automated training pipeline, continuous deployment, automatic scaling, and so on.
How will the generative AI model be updated and retrained as new data becomes available?	Through periodic retraining, incremental learning, transfer learning, and so on.
What are the requirements for automating monitoring and management?	Automated alerts, automatic scaling, self-heal ing, and so on.
What is your preferred deployment environme nt for generative AI workloads?	A hybrid approach that uses AWS for model training and our on-premises infrastructure for inference to meet data residency requireme nts.
Are there any specific cloud platforms you prefer for generative AI deployments?	AWS services, particularly Amazon SageMaker AI for model development and deployment, and Amazon Bedrock for foundation models.
What containerization technologies are you considering for generative AI workloads?	We want to standardize on Docker container s that are orchestrated with Kubernetes to

Question	Example response
	ensure portability and scalability across our hybrid environment.
Do you have any preferred tools for CI/CD in your generative AI pipeline?	GitLab for version control and CI/CD pipelines, integrated with Jenkins for automated testing and deployment.
What orchestration tools are you considering for managing generative AI workflows?	Apache Airflow for workflow orchestration, particularly for data preprocessing and model training pipelines.
Do you have any specific requirements for on-premises infrastructure to support generative AI workloads?	We're investing in GPU-accelerated servers and high-speed networking to support on-premises inference workloads.
How do you plan to manage model versionin g and deployment across different environme nts?	We plan to use MLflow for model tracking and versioning, and integrate it with our Kubernetes infrastructure for seamless deployment across environments.
What monitoring and observability tools are you considering for generative AI deploymen ts?	Prometheus for metrics collection and Grafana for visualization, with additional custom logging solutions for model-specific monitorin g.
How are you addressing data movement and synchronization in a hybrid deployment model?	We will use AWS DataSync for efficient data transfer between on-premises storage and AWS, with automated synchronization jobs that are scheduled based on our training cycles.
What security measures are you implementing for generative AI deployments across different environments?	We will use IAM for cloud resources, integrate d with our on-premises Active Directory to implement end-to-end encryption and network segmentation to secure data flows.

Data strategy

Question	Example response
What specific data types are crucial for your generative AI workloads, and what percentage of these are currently accessible?	Customer call logs and product reviews data are crucial. Currently, 85% of these data types are accessible for our generative AI projects.
How do you ensure and measure the quality of your data?	We have implemented data quality metrics, including completeness, accuracy, consisten cy, and timeliness. We use automated tools to regularly assess these metrics and have a dedicated team for data cleansing and enrichment.
What percentage of your data meets your quality standards for generative AI use?	Currently, 78% of our data meets our quality standards. We're aiming for 95% within the next 12 months through improved data cleaning processes.
How do you plan to build trust about data usage in generative AI among your stakehold ers?	We're implementing an AI ethics board, providing clear explanations of AI decisions, and conducting quarterly AI audits to ensure transparency and fairness.
How comprehensive is your documentation for data sources and lineage?	We maintain a detailed data catalog that includes metadata for all our data sources, including origin, update frequency, and usage. We use data lineage tools to track how data flows and transforms across our systems.
How do you ensure diversity in your datasets to prevent bias in AI models?	We actively source data from diverse demographics and regularly audit our datasets for representational bias. We also use synthetic data generation techniques to balance under-represented categories.

Question	Example response
What is your data refresh rate for critical generative AI models, and how do you determine this frequency?	Critical models are refreshed weekly. This frequency is determined by A/B testing performance metrics, and we aim for no more than 2% degradation between refreshes.
How many versions of critical datasets do you maintain and for how long?	We maintain the last five versions of each critical dataset, with a retention period of 18 months for each version.
How many cross-functional teams are involved in your generative AI initiatives and have access to your data?	We have three cross-functional teams. Each team includes data scientists, domain experts, ethicists, and business analysts.
What data governance policies and practices do you have in place?	We have a cross-functional data governanc e committee that oversees our data policies. We've implemented role-based access controls, data classification schemes, and regular audits to ensure compliance with our governance framework.
What measures do you have in place to ensure data privacy, obtain proper consent, and maintain confidentiality?	We have implemented a comprehensive data privacy framework aligned with GDPR and CCPA. This includes obtaining explicit consent for data usage, implementing data anonymiza tion techniques, and regular privacy impact assessments.
What percentage of your AI training datasets have been audited for bias in the last quarter?	70% of our AI training datasets were audited for bias last quarter. We're implementing automated bias detection tools to reach 100% quarterly audits.
What is your current data processing capacity, and how much do you project needing for future generative AI workloads?	Our current capacity is 10 TB/day. We project needing 30 TB/day within a year and are scaling our infrastructure to meet this demand.

Question	Example response
What is your strategy for balancing data privacy with the data needs of generative AI models?	We're implementing advanced anonymization techniques and synthetic data generation. Our goal is to increase our usable data for AI by 40% while reducing privacy risks by 60% over the next year.
What percentage of your machine learning (ML) datasets are accurately labeled, and what's your target accuracy rate?	Currently, 85% of our ML datasets are accurately labeled. We're targeting a 95% accuracy rate within the next quarter by employing both human and automated labeling techniques.

Translating assessment insights into actionable outcomes

This section provides a framework for analyzing the questionnaire responses and using those insights to shape the target architecture and other key deliverables of the generative AI modernization initiative. This framework bridges the gap between data collection and implementation, and ensures that the assessment directly informs and drives your modernization strategy.

Target architecture definition:

- Use the questionnaire responses to inform the selection of cloud services and design of data pipelines.
- Make sure that the architecture design supports scalability and interoperability as highlighted in the guide.

Customer readiness evaluation:

- Analyze the questionnaire responses related to current infrastructure, processes, and organizational culture.
- Identify gaps and create a plan to address them. Prioritize gaps that are critical for MVP success.

Use case and stretch goals:

- Extract specific business problems from the questionnaire responses to define clear use case goals.
- Set stretch goals that align with your organization's long-term vision for generative AI modernization.

Effort estimation:

- Use the questionnaire data to estimate resources, time, and budget for both the MVP and full implementation.
- Create a phased approach that starts with the MVP, and outline subsequent phases.

Enablement needs:

- Based on the questionnaire responses, identify skill gaps and training needs.
- Develop a training plan that supports both immediate MVP needs and long-term generative AI adoption.

Implementation plan:

- Create a comprehensive roadmap that starts with the MVP and outlines steps toward full generative AI modernization.
- Define clear milestones and deliverables for each phase of the implementation.

Practical steps:

- **Prioritization matrix**: Create a matrix that maps questionnaire responses to the six outcomes to help prioritize features and efforts.
- **Iterative approach**: Design the MVP to be the first iteration in a series of planned releases, where each release builds toward the full target architecture.
- **Stakeholder alignment**: Use the questionnaire results to align stakeholders on MVP scope and the phased approach to achieving all outcomes.
- **Continuous feedback loop**: Implement mechanisms to gather feedback after MVP deployment, and use insights to refine plans for subsequent phases.
- **Agile implementation**: Adopt an agile methodology that allows for flexibility in addressing all outcomes over time, starting with the most critical outcomes in the MVP.

Next steps

After you complete the generative AI workload assessment, follow these steps:

1. **Deliver a detailed target architecture**
   - Objective: The solution architect creates a comprehensive target architecture that aligns with the organization's goals and the outcomes of the assessment.
   - Components: This architecture includes the design of data ingestion, integration points, and system interoperability to ensure scalability, reliability, and performance optimization.
2. **Explain how specific AWS services fit the use case**
   - Service mapping: Identify and map specific AWS services that best fit the identified use cases.
   - Benefits: Highlight how these services address specific business needs, enhance efficiency, and provide scalability.
3. **Provide optional alternative solutions with pros and cons**
   - Alternatives: Present alternative solutions that could also meet the organization's requirements.
   - Analysis: Offer a detailed analysis of the advantages and disadvantages of each alternative by considering factors such as cost, complexity, and alignment with business goals.
4. **Provide detailed price estimation of AWS services**
   - Cost analysis: Deliver a detailed cost estimation for the proposed AWS services, including potential usage scenarios and pricing models.
   - Budget alignment: Make sure that the cost aligns with the organization's budgetary constraints and provides a clear understanding of the financial implications.
5. **Get feedback on the proposed architecture**
   - Stakeholder engagement: Engage with stakeholders to present the proposed architecture and gather feedback.
   - Iterative improvement: Use the feedback to refine and improve the solution, and confirm that it meets the needs and expectations of all stakeholders.

FAQ

What is the primary objective of the generative AI workload assessment?

The primary objective of the assessment is to evaluate an organization's readiness for modernizing their generative AI workloads, identify use cases, and develop a target solution architecture. It aims to define modernization requirements, determine implementation scope, and prepare for successful generative AI modernization.

Who should use this assessment?

This assessment is for solutions architects, enterprise architects, and application architects who want to assess the technical aspects of generative AI modernization. It is also useful for program managers and people managers to gauge overall readiness, resource allocation, and enablement needs.

What are the key components evaluated in the assessment?

The assessment covers overall readiness, use case, architecture, storage, regulations and compliance, integration, testing, deployment automation, and data strategy. These components are crucial for determining the technical and organizational readiness for generative AI modernization adoption.

How does the assessment help define the target architecture?

The assessment provides a structured approach to evaluate current systems and identify improvements. It helps you select appropriate technologies and design scalable architectures that align with business goals and use case requirements.

What are the benefits of conducting a generative AI workload assessment?

Benefits include enhanced efficiency, improved decision-making, compliance assurance, innovation fostering, and scalability preparation. The assessment establishes a strategic approach to generative AI modernization, and maximizes potential benefits while mitigating risks.

How can organizations ensure successful implementation following the assessment?

Organizations should develop a clear implementation plan that includes defined milestones, engage stakeholders early, and adopt an iterative approach. Establishing a Center of Excellence (CoE) and focusing on talent development are also recommended best practices.

What challenges might organizations face during the assessment?

Challenges might include resistance to change, data quality issues, and compliance complexities. Addressing these challenges requires fostering a culture of innovation, ensuring data readiness, and implementing robust security measures.

How does the assessment address regulatory and compliance requirements?

The assessment evaluates current compliance measures and identifies gaps. It ensures that target solutions adhere to relevant regulations and data privacy laws, and incorporate security best practices to protect sensitive information.

What role does stakeholder engagement play in the assessment process?

Stakeholder engagement is crucial for gaining buy-in, aligning modernization initiatives with business objectives, and ensuring successful implementation. Early involvement and clear communication of benefits are key to overcoming resistance and fostering support.

How can organizations measure the success of their generative AI modernization initiatives after the assessment?

Success can be measured by using key performance indicators (KPIs) that align with business goals. Regular monitoring and evaluation of these metrics help guide decision-making and demonstrate the value of generative AI modernization to stakeholders.

How does the assessment approach differ for organizations of varying sizes (small, medium, or enterprise) or industries?

Small organizations:

- Might have limited resources and expertise for comprehensive assessments
- Likely to focus on specific high-impact use cases instead of enterprise-wide adoption
- Might rely more heavily on third-party tools and services for assessment
- Assessment process might be less formal and more agile

Mid-sized organizations:

- Often have dedicated IT or data teams but might lack specialized AI expertise
- Might take a phased approach, starting with pilot projects in key departments
- Need to balance innovation with existing systems and processes
- Assessment likely involves cross-functional teams

Enterprise organizations:

- Typically have dedicated AI/ML teams and more resources for comprehensive assessment
- Need to consider complex integrations with existing enterprise systems
- Might have industry-specific regulatory requirements to factor in
- Assessment often involves formal governance processes

Resources

- Generative AI on AWS
- AWS offers new artificial intelligence, machine learning, and generative AI guides to plan your AI strategy (AWS blog post)
- Best practices to build generative AI applications on AWS (AWS blog post)
- Generative AI Application Builder on AWS (AWS Solutions Library)
- Generative AI capabilities (*AWS Security Reference Architecture*)
- AWS generative AI best practices framework (*AWS Audit Manager User Guide*)
- Choosing a generative AI service (AWS decision guide)
- What is Amazon Bedrock? (*Amazon Bedrock User Guide*)
- What is Amazon SageMaker AI? (*Amazon SageMaker AI Developer Guide*)

Document history

The following table describes significant changes to this guide. If you want to be notified about future updates, you can subscribe to an RSS feed.

Change	Description	Date
Initial publication	—	November 6, 2024