



Évaluation de la charge de travail générative de

AWS Conseils prescriptifs



AWS Conseils prescriptifs: Évaluation de la charge de travail générative de

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

Les marques et la présentation commerciale d'Amazon ne peuvent être utilisées en relation avec un produit ou un service qui n'est pas d'Amazon, d'une manière susceptible de créer une confusion parmi les clients, ou d'une manière qui dénigre ou discrédite Amazon. Toutes les autres marques commerciales qui ne sont pas la propriété d'Amazon appartiennent à leurs propriétaires respectifs, qui peuvent ou non être affiliés ou connectés à Amazon, ou sponsorisés par Amazon.

Table of Contents

Introduction	1
Objectif de ce guide	2
Public cible et avantages	2
Portée	3
Résultats commerciaux ciblés	4
Considérations relatives à l'évaluation et conditions préalables	7
Commencez par des cas d'utilisation clairs	7
Garantir l'alignement des activités	8
Mettre en œuvre la gouvernance et la supervision	8
Données d'adresse et prérequis techniques	8
Tenez compte des besoins en ressources informatiques	9
Aborder les implications en matière de confidentialité et de sécurité	9
Impliquer les parties prenantes dès	9
Itérer et apprendre	9
Questionnaire d'évaluation de la charge de travail de	10
État de préparation	11
Cas d'utilisation	13
Architecture	16
Stockage	17
Réglementation et conformité	18
Intégration	19
Test	22
Déploiement et automatisation	23
Stratégie en matière de données	26
Traduire les informations issues des évaluations en résultats exploitables	29
Étapes suivantes	31
FAQ	32
Quel est l'objectif principal ?	32
Qui devrait utiliser cette évaluation ?	32
Quels en sont les éléments clés ?	32
Comment cela aide-t-il à définir l'architecture ?	32
Quels en sont les avantages ?	33
Comment pouvons-nous le mettre en œuvre avec succès ?	33
Quels sont les défis ?	33

Quelles sont les exigences réglementaires et de conformité ?	33
Quel est le rôle des parties prenantes ?	33
Comment mesurer le succès ?	34
En quoi l'approche diffère-t-elle en fonction de la taille de l'organisation ?	34
Ressources	36
Historique du document	37
.....	xxxviii

Évaluation de la charge de travail générative de

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

Novembre 2024 ([historique du document](#))

L'évaluation de la charge de travail d'IA générative est une méthode stratégique visant à évaluer et à améliorer le niveau de préparation d'une organisation à créer ou à mettre à jour ses charges de travail d'IA générative. Cette évaluation est importante car l'intégration de l'IA générative dans les opérations commerciales peut modifier considérablement la façon dont les choses fonctionnent et peut apporter de nouvelles performances et capacités. Cependant, pour réussir à adopter l'IA générative, il est essentiel de bien comprendre les systèmes actuels et d'avoir un plan clair pour le futur.

Les charges de travail génératives liées à l'IA font référence à des tâches informatiques impliquant l'utilisation de modèles d'intelligence artificielle capables de créer de nouveaux contenus, tels que du texte, des images, du code ou d'autres types de données. Ces charges de travail nécessitent généralement une puissance de calcul importante, du matériel spécialisé tel que GPUs, et de grands ensembles de données pour la formation et l'inférence. L'intégration des charges de travail génératives liées à l'IA dans les opérations présente plusieurs défis :

- Exigences en matière d'infrastructure : provisionnement des ressources informatiques importantes et du matériel spécialisé requis par les modèles d'IA générative.
- Gestion des données : garantir la qualité, la confidentialité et la conformité des données lors de la gestion de grands ensembles de données.
- Manque de compétences : manque d'expertise dans les technologies d'intelligence artificielle et le déploiement de modèles.
- Considérations éthiques : remédier aux préjugés, à l'équité et à la transparence dans le contenu généré par l'IA.
- Complexité de l'intégration : intégration fluide de l'IA générative dans les flux de travail existants et les systèmes existants.
- Gestion des coûts : équilibre entre les avantages potentiels et les coûts élevés de mise en œuvre et d'exploitation.

Pour surmonter ces défis, il faut une planification minutieuse, des investissements dans les infrastructures et les talents, ainsi qu'une approche stratégique de la mise en œuvre.

Objectif de ce guide

L'IA générative devient rapidement un élément essentiel dans de nombreux secteurs. Elle offre des opportunités de transformation mais pose également des défis en termes d'intégration, de conformité et d'évolutivité. De nombreuses organisations ont du mal à tirer pleinement parti de l'IA en raison de la faiblesse des bases technologiques, de la résistance au changement et des problèmes de qualité des données. L'évaluation de la charge de travail de l'IA générative permet de relever ces défis en identifiant les exigences de modernisation, en définissant le champ de mise en œuvre et en remettant en question les systèmes et les modes de pensée existants. Il aide également à déterminer le minimum de produits viables (MVPs) et vous aide à développer une architecture de solution cible, garantissant ainsi une approche structurée et stratégique de l'adoption de l'IA.

Ce guide constitue une approche structurée pour aider les organisations à faire face aux complexités liées à l'adoption de technologies d'IA générative. Au lieu de définir clairement les exigences dès le départ, le guide aide à :

- Identifier les cas d'utilisation potentiels de l'IA générative au sein de votre organisation.
- Évaluer l'état de préparation de votre entreprise à l'adoption de l'IA générative.
- Définition et affinement des objectifs des cas d'utilisation et des objectifs ambitieux.
- Déterminer la portée et les exigences de la mise en œuvre de l'IA générative.
- Développement d'une architecture de solution cible.

Public cible et avantages

Cette évaluation est spécialement conçue pour les architectes de solutions, les architectes d'entreprise et les architectes d'applications qui souhaitent évaluer les aspects techniques de la modernisation de la charge de travail générative de l'IA. Il est également utile pour les responsables des programmes et des personnes qui souhaitent évaluer l'état de préparation global de leur équipe, l'allocation des ressources et les exigences en matière d'habilitation. Les meilleures pratiques du secteur soulignent l'importance d'une évaluation complète pour garantir l'état de préparation à l'adoption de l'IA. Cela inclut l'évaluation de l'architecture, du stockage, de la conformité, de l'intégration, des tests, du déploiement et de l'automatisation.

Portée

Les sujets suivants sont concernés par la méthode d'évaluation de la charge de travail de l'IA générative :

- Technologies et modèles actuels d'IA générative (par exemple, grands modèles de langage, modèles de génération d'images)
- Applications d'IA restreintes qui utilisent des techniques génératives
- Intégration de l'IA générative aux systèmes et flux de travail existants
- Stratégies de données pour la formation et le réglage précis des modèles d'IA générative
- Considérations éthiques et pratiques responsables en matière d'IA pour les applications d'IA générative actuelles
- Stratégies de test et de déploiement pour l'IA générative dans les environnements de production
- Considérations relatives à la sécurité et à la confidentialité pour les implémentations d'IA générative
- Optimisation des performances et évolutivité des charges de travail génératives liées à l'IA
- Cas d'utilisation et applications de l'IA générative dans divers secteurs
- Évaluation des résultats de l'IA générative et des processus d'assurance qualité

Les sujets suivants sont hors de portée :

- Scénarios d'intelligence générale artificielle (AGI) et de superintelligence artificielle (ASI)
- Les futures avancées spéculatives de l'IA au-delà des modèles génératifs actuels
- Applications de l'informatique quantique dans le domaine de l'IA
- Informatique neuromorphique et interfaces cerveau-ordinateur
- Conscience et conscience de soi dans les systèmes d'IA
- Impacts sociétaux à long terme de l'IA avancée au-delà des applications d'IA générative actuelles
- Cadres réglementaires pour les futures technologies d'IA hypothétiques
- Débats philosophiques sur la nature de l'intelligence et de la conscience dans les machines
- Cas extrêmes ou cas d'utilisation hautement spéculatifs de l'IA
- Spécifications techniques détaillées des modèles ou architectures d'IA propriétaires

Résultats commerciaux ciblés

L'évaluation de la charge de travail de l'IA générative vise à obtenir plusieurs résultats ciblés qui sont essentiels pour moderniser avec succès les charges de travail de l'IA générative. Ces résultats garantissent que les organisations sont bien préparées à intégrer les technologies d'IA de manière efficace et efficiente.

Pour chaque résultat visé, l'évaluation de la charge de travail de l'IA générative se concentre sur :

- **Interdépendances** : Identifiez et clarifiez les interdépendances entre le résultat et les autres aspects du processus de modernisation. Cela implique de comprendre comment un résultat peut influencer ou être influencé par d'autres, afin de garantir une approche holistique de la modernisation.
- **Harmonisation des parties prenantes** : Décrivez des stratégies pour aligner les différentes parties prenantes sur chaque résultat. Cela implique de communiquer la valeur et l'impact de chaque résultat aux différents niveaux organisationnels et départements, afin de favoriser l'adhésion et le soutien.
- **Hiérarchisation** : dans les cas où plusieurs cas d'utilisation ou résultats sont identifiés, fournissez un cadre pour les hiérarchiser en fonction de facteurs tels que l'impact commercial, les besoins en ressources et l'alignement stratégique.
- **Amélioration continue** : pour chaque résultat, établir des mécanismes d'évaluation et de perfectionnement continus. Cela garantit que les efforts de modernisation restent adaptatifs et réactifs à l'évolution des paysages technologiques et des besoins commerciaux.

Voici une discussion détaillée de chaque résultat visé :

Architecture cible

- **Définition** : L'évaluation permet de définir une architecture cible claire et évolutive pour les charges de travail génératives liées à l'IA.
- **Composants** : Cela inclut la sélection des services cloud appropriés, la conception de pipelines de données et la garantie de l'interopérabilité du système.
- **Avantages** : Une architecture bien définie favorise l'évolutivité, la fiabilité et l'optimisation des performances, et fournit une base solide pour la modernisation.

Préparation du client

- **Évaluation** : Évaluez l'état actuel de l'infrastructure, des processus et de la culture de l'organisation afin de déterminer si elle est prête à adopter la modernisation de l'IA générative.
- **Critères** : Cela implique d'évaluer les capacités techniques, la qualité des données et la volonté organisationnelle d'accepter le changement.
- **Résultat** : L'identification des lacunes et des domaines à améliorer garantit que l'organisation est prête à effectuer une transition harmonieuse vers des solutions et des technologies modernes.

Utilisez les objectifs du cas et étirez les objectifs

- Les objectifs du cas d'utilisation établissent des objectifs clairs pour la mise en œuvre de la solution cible, en se concentrant sur des problèmes ou des opportunités commerciaux spécifiques.

Un objectif de cas d'utilisation dans le contexte de la modernisation de l'IA générative fait référence à un objectif spécifique et mesurable qu'une organisation vise à atteindre en mettant en œuvre des solutions d'IA générative. Ces objectifs sont généralement alignés sur des objectifs commerciaux plus généraux et visent à relever des défis ou à saisir des opportunités spécifiques au sein de l'organisation. Voici des exemples d'objectifs de cas d'utilisation :

- Réduire le temps de réponse du service client de 50 % en utilisant des chatbots génératifs alimentés par l'IA.
- Améliorer l'efficacité de la révision du code de 30 % grâce à une analyse de code générative assistée par l'IA.
- Améliorer la précision de la détection des fraudes de 25 % grâce à la reconnaissance générative des formes par IA.
- Les objectifs ambitieux définissent des cibles ambitieuses qui repoussent les limites de ce que la modernisation de l'IA générative peut atteindre au sein de l'organisation.
- **Impact** : La définition d'objectifs à la fois réalisables et ambitieux permet d'aligner les initiatives de modernisation de l'IA générative sur les objectifs commerciaux stratégiques et d'encourager l'innovation.

Estimation de l'effort

- **Objectif** : Une estimation précise des efforts facilite la planification des ressources et garantit que les projets sont livrés dans les délais et dans les limites du budget.
- **Portée** : Estimez les ressources, le temps et le budget nécessaires pour mettre en œuvre le plan de modernisation de l'IA générative.

- Facteurs : Tenez compte de la complexité technique, des défis d'intégration et des risques potentiels.

Besoins en matière d'habilitation

- Formation et développement : identifiez les compétences et les connaissances requises pour une adoption réussie de l'IA générative.
- Ressources : Déterminer le besoin de programmes de formation, d'ateliers et d'autres activités d'habilitation.
- Résultat : Veiller à ce que le personnel possède les compétences nécessaires améliore l'efficacité des initiatives de modernisation de l'IA générative et favorise le succès à long terme.

Plan de mise en œuvre

- Feuille de route : Élaborez un plan détaillé décrivant les étapes nécessaires à la modernisation de l'IA générative.
- Jalons : définissez les principaux jalons et livrables pour suivre les progrès.
- Avantages : Un plan de mise en œuvre clair fournit une orientation et une responsabilisation, et facilite une approche structurée de la modernisation de l'IA générative.

Considérations relatives à l'évaluation et conditions préalables

Commencez par des cas d'utilisation clairs

Identifiez les problèmes ou opportunités commerciaux spécifiques auxquels l'IA générative peut répondre. Concentrez-vous sur les cas d'utilisation qui correspondent aux objectifs commerciaux stratégiques et offrent des avantages mesurables. Priorisez les cas d'utilisation qui ciblent les défis courants au sein de l'organisation afin de garantir que l'architecture de la solution puisse servir de modèle pour plusieurs scénarios.

Lancer le processus d'évaluation avec une compréhension générale des applications potentielles de l'IA générative est bénéfique mais pas obligatoire. Le [questionnaire](#) inclus dans ce guide répond à différents niveaux de préparation, qu'il s'agisse d'organisations dont les cas d'utilisation sont bien définis ou de celles qui n'ont que des idées générales. Le processus d'évaluation sert à :

- Affinez et clarifiez ces idées de cas d'utilisation initiaux.
- Identifiez de nouveaux cas d'utilisation potentiels.
- Définissez des objectifs spécifiques et mesurables pour chaque cas d'utilisation.
- Évaluez la faisabilité et l'impact potentiel de chaque cas d'utilisation.

Prenons un exemple hypothétique : une société de services financiers décide d'explorer la modernisation de l'IA générative. Ils commencent par une idée générale de l'amélioration de leur service client et de leurs processus de détection des fraudes.

- Évaluation initiale : Le questionnaire les aide à évaluer leurs systèmes actuels, la qualité des données et l'état de préparation de l'organisation à l'adoption de l'IA générative.
- Affinement des cas d'utilisation : grâce au processus d'évaluation, ils affinent leurs idées initiales en deux cas d'utilisation spécifiques :
 - Implémentation d'un chatbot génératif alimenté par l'IA pour les demandes des clients
 - Utilisation de l'IA générative pour détecter les fraudes transactionnelles en temps réel
- Définition des objectifs : pour chaque cas d'utilisation, ils définissent des objectifs spécifiques :
 - Réduire le temps de réponse du service client de 40 % en 6 mois

- Améliorez la précision de la détection des fraudes de 20 % et réduisez les faux positifs de 15 %
- Des objectifs ambitieux : ils ont également fixé ces objectifs ambitieux :
 - Obtenez 80 % de satisfaction client grâce aux réponses assistées par l'IA
 - Développez un modèle de détection prédictive des fraudes qui identifie les nouveaux modèles de fraude
- Définition du MVP : le questionnaire les aide à déterminer un MVP pour chaque cas d'utilisation, en se concentrant sur les fonctionnalités essentielles qui apportent une valeur immédiate.
- Architecture cible : Enfin, ils développent une architecture cible qui prend en charge un ou les deux cas d'utilisation et garantit l'évolutivité et l'intégration avec les systèmes existants.

Garantir l'alignement des activités

Alignez les initiatives d'IA générative sur la stratégie et les objectifs commerciaux globaux. Pour chaque cas d'utilisation, élaborer une proposition de valeur claire qui montre comment l'IA générative contribue à la croissance, à l'efficacité ou à l'innovation des entreprises. Établissez des mesures pour mesurer l'impact des mises en œuvre de l'IA générative sur les indicateurs de performance clés (KPIs).

Mettre en œuvre la gouvernance et la supervision

Créez un comité directeur interfonctionnel chargé de superviser les initiatives en matière d'IA générative. Élaborer des politiques et des directives pour une utilisation responsable de l'IA, en tenant compte des considérations éthiques et des biais potentiels. Établissez un processus d'examen des projets d'IA générative afin de garantir la conformité aux normes organisationnelles et aux exigences réglementaires.

Données d'adresse et prérequis techniques

Évaluez et améliorez la qualité des données, et mettez en œuvre des pratiques de gouvernance des données afin de garantir des entrées fiables pour les modèles d'IA génératifs. Développez une stratégie de données qui aborde la collecte, le stockage et la gestion des données spécifiques aux besoins en IA générative. Évaluez et améliorez l'infrastructure de données pour prendre en charge le volume et la rapidité des données nécessaires aux charges de travail génératives liées à l'IA.

Tenez compte des besoins en ressources informatiques

Évaluez l'infrastructure informatique actuelle et identifiez les lacunes en matière de capacité de calcul pour les charges de travail génératives liées à l'IA. Planifiez des ressources informatiques évolutives, en envisageant des options telles que les services cloud ou les clusters de calcul hautes performances sur site. Optimisez l'allocation des ressources pour équilibrer les performances et la rentabilité des charges de travail de formation et d'inférence.

Aborder les implications en matière de confidentialité et de sécurité

Mettez en œuvre des mesures de sécurité robustes pour protéger les données sensibles utilisées dans le cadre de la formation et des opérations liées à l'IA générative. Garantisiez le respect des réglementations en matière de protection des données telles que le règlement général sur la protection des données (RGPD) ou la loi californienne sur la protection de la vie privée des consommateurs (CCPA) lors du traitement des informations personnelles. Développez des protocoles pour le déploiement et la surveillance sécurisés des modèles afin d'empêcher tout accès non autorisé ou toute utilisation abusive des capacités d'IA générative.

Impliquer les parties prenantes dès

Impliquez les principales parties prenantes dès le début pour obtenir l'adhésion et le soutien des dirigeants. Communiquez clairement les avantages et l'impact potentiel des initiatives de modernisation, en particulier pour les charges de travail génératives liées à l'IA. Fournir des formations et des ressources pour aider les parties prenantes à comprendre les technologies d'IA générative et leurs implications.

Itérer et apprendre

Adoptez une approche progressive qui vous permet d'affiner les solutions cibles. Utilisez des boucles de rétroaction pour améliorer en permanence l'architecture et les processus de charge de travail. Évaluez régulièrement les performances et l'impact des mises en œuvre de l'IA générative, et ajustez les stratégies selon les besoins en fonction des résultats concrets et de l'évolution des besoins de l'entreprise.

Questionnaire d'évaluation de la charge de travail de

Les sections suivantes proposent des questions que vous pouvez utiliser pour évaluer les différents aspects de la modernisation des charges de travail génératives liées à l'IA pour votre organisation. Ce questionnaire complet évalue le degré de préparation de votre entreprise à adopter et à mettre en œuvre des charges de travail basées sur l'IA générative en posant des questions portant sur des domaines clés, notamment les cas d'utilisation, l'architecture, le stockage, la conformité, l'intégration, les tests, le déploiement et la stratégie en matière de données. En abordant les aspects critiques de la mise en œuvre de l'IA générative, de l'infrastructure technique aux considérations réglementaires, ce questionnaire vous aide à identifier les forces, les lacunes et les opportunités de votre parcours de modernisation de l'IA.

Rubriques :

- [État de préparation](#)
- [Cas d'utilisation](#)
- [Architecture](#)
- [Stockage](#)
- [Réglementation et conformité](#)
- [Intégration](#)
- [Test](#)
- [Déploiement et automatisation](#)
- [Stratégie en matière de données](#)

Vous pouvez également télécharger le questionnaire au format Microsoft Excel et l'utiliser pour enregistrer vos informations.

 [le questionnaire](#)

Téléch

État de préparation

Question	Exemple de réponse
Disposez-vous de AWS comptes qui peuvent être utilisés pour ces charges de travail ?	Oui ou non
Avez-vous déjà conclu un accord d'entreprise avec AWS ?	Oui ou non
Dans quelle mesure votre infrastructure cloud actuelle est-elle évolutive pour gérer les charges de travail génératives liées à l'IA ?	Notre infrastructure cloud est hautement évolutive, avec des capacités de mise à l'échelle automatique pour les ressources informatiques et les systèmes de stockage distribués conçus pour gérer efficacement les charges de travail d'IA générative à grande échelle.
Disposez-vous de capacités de pipeline de données pour le prétraitement et l'ingénierie des fonctionnalités à grande échelle ?	Nos pipelines de données utilisent des frameworks de traitement distribués tels qu'Apache Spark pour le prétraitement des données à grande échelle et l'ingénierie des fonctionnalités, avec un support pour le traitement des données par lots et en streaming .
Disposez-vous de capacités de provisionnement et de gestion de comptes ?	Oui ou non
Comment décririez-vous les connaissances de votre organisation en matière d'IA et sa volonté d'adopter les technologies d'IA générative ?	Notre organisation a beaucoup investi dans des programmes de formation en IA, et la plupart du personnel technique a suivi une formation de base sur l'IA et le ML. L'organisation possède une culture de l'innovation qui intègre les nouvelles technologies, notamment l'IA générative.

Question	Exemple de réponse
Quelle est l'expertise en matière d'intelligence artificielle et de machine learning existant au sein de votre organisation, et comment est-elle distribuée ?	Nous disposons d'un centre d'excellence dédié à l'IA composé de data scientists et d'ingénieurs ML expérimentés. Nous améliorons les compétences des experts du domaine de différentes unités commerciales afin qu'ils maîtrisent l'IA et identifient les cas d'utilisation de l'IA générative.
Avez-vous une analyse de rentabilisation de haut niveau qui expose les objectifs, les avantages et les coûts du programme cloud ?	Oui ou non
Quel est votre calendrier pour mettre la solution en production ?	Des semaines, des mois, etc.
Un engagement financier a-t-il été pris par vos principales parties prenantes (par exemple, CFO, CIT/CTO, COO) ?	Oui ou non
Comment garantisiez-vous le respect des réglementations en matière de protection des données dans le cadre de vos initiatives d'IA générative ?	Nous disposons d'une équipe dédiée à la conformité qui travaille en étroite collaboration avec nos équipes d'IA. Nous réalisons régulièrement des évaluations de l'impact sur la vie privée, mettons en œuvre des principes de protection des données dès la conception et conservons des dossiers de traitement des données détaillés pour tous les projets d'IA générative.
Quel est le degré de maturité de vos systèmes existants qui s'intègrent aux nouvelles technologies d'IA générative ?	Notre architecture informatique est basée sur des microservices et APIs permet une intégration flexible des nouvelles technologies génératives d'IA. Ces systèmes sont normalisés selon des formats de données et des protocoles communs afin de garantir l'interopérabilité.

Question	Exemple de réponse
Quelle expérience avez-vous en matière d'opérationnalisation de modèles de machine learning, et comment cela pourrait-il s'appliquer aux systèmes d'IA générative ?	Nous avons établi MLOps des pratiques, notamment des pipelines de déploiement de modèles automatisés, des systèmes de surveillance et des cadres de tests A/B. Ces pratiques sont en cours d'adaptation pour répondre aux exigences uniques des modèles d'IA générative à grande échelle.

Cas d'utilisation

Question	Exemple de réponse
Quel est l'objectif principal ou le critère de réussite du cas d'utilisation ?	Pour améliorer le temps de réponse du support client, augmenter le taux de conversion des ventes, améliorer les recommandations de produits. Également : pour améliorer la satisfaction des utilisateurs, le taux d'achèvement des tâches, la qualité des réponses, etc.
Comment ce cas d'utilisation s'aligne-t-il sur les objectifs stratégiques de votre organisation ?	Cela correspond à notre objectif stratégique d'améliorer la satisfaction des clients en réduisant les temps de réponse du service client.
Quel est le volume de données ou de demandes attendu pour le cas d'utilisation ?	500 transactions par seconde (TPS).
Quels types de sources de données sont nécessaires pour prendre en charge vos charges de travail génératives liées à l'IA ?	Bases de données structurées internes (dossiers clients, données de vente, etc.) ; données textuelles non structurées provenant de documents, d'e-mails et de réseaux sociaux ; fichiers audio et vidéo pour les tâches de reconnaissance vocale et d'image ; données de streaming en temps réel provenant d'appareils

Question	Exemple de réponse
À quelle fréquence devez-vous mettre à jour ou actualiser les données provenant de ces sources ?	Is et de capteurs IoT ; ensembles de données publics et à des APIs fins d'enrichissement. Bases de données transactionnelles : mises à jour en temps quasi réel ; référentiels de documents : mises à jour quotidiennes par lots ; flux de réseaux sociaux : mises à jour horaires ; données des capteurs IoT : diffusion continue en temps réel ; ensembles de données publics : mises à jour mensuelles ou trimestrielles.
Quels formats de données vos modèles d'IA générative ont-ils besoin en entrée ?	Données structurées : tables de base de données CSV, JSON et SQL ; données texte : texte brut, PDF et HTML ; données d'image : JPEG, PNG et TIFF ; données audio : WAV et MP3 ; données vidéo : MP4 et AVI.
Quelles sont vos principales préoccupations en matière de qualité des données pour les charges de travail génératives liées à l'IA ?	Exhaustivité : garantir qu'aucun champ critique n'est manquant ; exactitude : vérification de l'exactitude des données et élimination des erreurs ; cohérence : maintien de formats et de valeurs uniformes entre les sources ; actualité : garantie que les données sont à jour pour une inférence en temps réel ; pertinence : confirmation que les données correspondent à la tâche spécifique d'IA générative.
Quelles sont les principales exigences de performance (par exemple, temps de réponse, débit, précision) ?	Précision de 95 % ; temps de réponse inférieur à 500 ms ; capacité à traiter 1 000 requêtes/sec. Haute précision (95% +), précision modérée (80-90%), meilleur effort, etc.
En avez-vous un autre KPIs pour mesurer le succès de ce cas d'utilisation ?	KPIs Les éléments clés incluent la réduction du taux d'erreur, le gain de temps par transaction et les scores de satisfaction client.

Question	Exemple de réponse
Quel est le niveau de précision du modèle souhaité, et quel est son équilibre avec le coût ?	Haute précision (> 90%) à coût modéré, précision modérée (70-80%) à faible coût, etc.
Quels sont les principaux cas d'utilisation ou scénarios de la solution d'IA générative ?	Chatbot du service client, génération de contenu, recommandation de produits, etc.
Quels sont les utilisateurs ou les personnalités cibles du système d'IA générative ?	Agents du service client, équipe marketing, employés, utilisateurs finaux, etc.
Quel est le volume de demandes ou d'utilisateurs attendu ?	1 000 demandes par jour ; 10 000 utilisateurs actifs par mois.
Existe-t-il des contraintes ou des exigences spécifiques liées aux cas d'utilisation ?	Réponse en temps réel, support multilingue, confidentialité des données, etc.
Disposez-vous d'un budget alloué au développement et à la maintenance de la solution d'IA générative ?	Le coût de développement initial est estimé à 200 000\$, avec des coûts de maintenance annuels de 50 000\$.
Quels sont le retour sur investissement (ROI) et la période de remboursement prévus pour ce cas d'utilisation ?	Retour sur investissement attendu de 150 % sur trois ans, avec une période de remboursement de 18 mois.
Y a-t-il des coûts cachés ou des économies potentielles à envisager ?	Les économies potentielles incluent la réduction des coûts des heures supplémentaires. Les coûts cachés peuvent impliquer une formation supplémentaire pour le personnel.
Quelles sont l'évolutivité et les futures possibilités d'extension de cette solution d'IA générative ?	La solution est conçue pour s'adapter à nos opérations, avec la possibilité de l'étendre à d'autres départements à l'avenir.
Comment garantir l'équité et atténuer les biais dans vos modèles d'IA générative ?	Nous prévoyons d'atténuer les biais grâce à une collecte de données diversifiée, à des audits réguliers des biais et à la mise en œuvre de techniques d'atténuation des biais.

Question	Exemple de réponse
Quels processus avez-vous mis en place pour répondre aux préoccupations éthiques ou aux conséquences imprévues ?	Nous gérons les préoccupations éthiques grâce à un plan établi de réponse aux incidents liés à l'IA, à des évaluations régulières des risques éthiques, à un système de signalement anonyme pour les employés, à la collaboration avec des experts externes en éthique, ainsi qu'à un suivi et à un ajustement continus des modèles déployés en fonction des commentaires.
Comment abordez-vous la hiérarchisation et le séquençage des évaluations de la charge de travail générative de l'IA dans les différents projets et départements de votre organisation ?	En menant une enquête de haut niveau auprès de tous les départements afin d'identifier les cas d'utilisation potentiels de l'IA générative et en les évaluant sur la base de trois critères clés : impact commercial, faisabilité technique et considérations éthiques. Les projets ayant un impact potentiel élevé, des obstacles techniques réduits et des préoccupations éthiques minimales sont prioritaires.

Architecture

Question	Exemple de réponse
Quel type de modèle ou d'architecture d'IA générative est envisagé ?	Transformateur, réseau neuronal convolutif (CNN), réseau neuronal récurrent (RNN), arbres de décision, etc.
Quelle est l'échelle ou le volume de données et de calculs attendus ?	Des millions d'utilisateurs, des pétaoctets de données, etc.
Quelles sont les exigences matérielles (par exemple, CPUs ou GPUs) pour la formation et l'inférence ?	Clusters de processeurs haut GPU de gamme, instances cloud, etc.

Question	Exemple de réponse
Comment le modèle d'IA générative sera-t-il mis à jour ou réentraîné au fil du temps ?	Grâce à un apprentissage continu, à une reconversion périodique, à des mises à jour manuelles, etc.
Quelles sont les exigences en matière de prétraitement des données et d'ingénierie des fonctionnalités ?	Nettoyage du texte, augmentation de l'image, sélection de fonctionnalités, etc.
Comment le système d'IA générative gèrera-t-il les cas extrêmes, les valeurs aberrantes ou les entrées peu fiables ?	En recourant à une supervision humaine, en demandant des éclaircissements, etc.
Quelles sont les exigences de latence pour l'application d'IA générative ?	Traitement par lots en temps réel, en temps quasi réel, etc.

Stockage

Question	Exemple de réponse
Où seront stockées les données d'entraînement ?	Dans le stockage dans le cloud (par exemple, Amazon S3, le stockage de fichiers, le stockage par blocs ou le stockage d'objets), dans le stockage sur site, etc.
Quelles sont les exigences de stockage pour les données de formation et les artefacts du modèle (par exemple, capacité, durabilité, disponibilité) ?	Stockage à l'échelle du pétaoctet, durabilité élevée (durabilité de 99,999999999 %), haute disponibilité, etc.
Quelles sont les exigences en matière de conservation et de sauvegarde des données de formation et des artefacts du modèle ?	Conservation des données pendant x ans, sauvegardes quotidiennes, sauvegardes hors site, etc.
Quels formats de fichiers sont principalement utilisés pour stocker vos ensembles	Fichiers Parquet pour les données structurées, HDF5 les grands tableaux multidimensionnels

Question	Exemple de réponse
de données de formation basés sur l'IA (par exemple, CSV, JSON, Parquet HDF5) ?	et les données non structurées telles que les images et le texte. Nous utilisons des formats spécialisés, par exemple TFRecord pour optimiser le chargement des données pendant l'entraînement.
Comment sont organisés vos ensembles de données d'entraînement : sous forme de fichiers individuels, dans des bases de données ou à l'aide de formats de données d'IA spécialisés ?	Les ensembles de données de petite à moyenne taille sont stockés sous forme de fichiers Parquet individuels dans le stockage d'objets pour plus de flexibilité. Les grands ensembles de données sont stockés dans une base de données distribuée (Cassandra) pour gérer l'échelle.
Utilisez-vous des techniques de compression ou d'encodage de données spécifiques aux données d'entraînement génératives issues de l'IA ?	Pour les données tabulaires, nous utilisons des techniques de codage par dictionnaire et de compression de bits disponibles dans Parquet. Pour les images, nous utilisons une compression JPEG avec perte avec des paramètres de qualité optimisés pour nos modèles.
Comment gérez-vous le versionnement et le stockage des différentes itérations d'ensembles de données de formation ? Quel impact cela a-t-il sur l'ensemble de vos besoins de stockage ?	Nous utilisons un système de version des données (DVC) intégré à notre plateforme ML.

Réglementation et conformité

Question	Exemple de réponse
Quelles sont les réglementations ou exigences de conformité pertinentes pour la solution d'IA générative (par exemple, RGPD, HIPAA, PCI-DSS) ?	RGPD pour le traitement des données personnelles, HIPAA pour les données de santé, PCI-DSS pour les données de paiement, etc.

Question	Exemple de réponse
Quels directives ou cadres éthiques en matière d'IA générative votre organisation a-t-elle adoptés ?	Nous avons mis en œuvre nos propres directives en matière d'IA responsable. Tous les projets d'IA générative font l'objet d'un examen éthique avant d'être approuvés et déployés.
Quelles sont les exigences de sécurité pour le système d'IA générative ?	Chiffrement des données, communication réseau sécurisée, audits de sécurité réguliers.
Quelles sont les exigences en matière de confidentialité et de protection des données ?	Anonymisation des données, chiffrement, contrôle d'accès, etc.
Quelles sont les exigences requises pour que la solution gère des données sensibles ou confidentielles ?	Contrôles d'accès stricts, masquage des données, exigences de résidence des données, etc.
Comment seront gérées l'authentification et l'autorisation des utilisateurs ?	En utilisant des clés d'API OAuth, une authentification unique (SSO) et un contrôle d'accès basé sur les rôles (RBAC).
Comment la solution sera-t-elle surveillée et gérée en production ?	En utilisant des outils de surveillance tels que Prometheus et Datadog, des outils de journalisation tels que ELK Stack, des systèmes d'alerte, etc.

Integration

Question	Exemple de réponse
Quelles sont les exigences pour intégrer la solution d'IA générative aux systèmes ou sources de données existants ?	REST APIs, files d'attente de messages, connecteurs de base de données, etc.

Question	Exemple de réponse
Comment les données seront-elles ingérées et prétraitées pour la solution d'IA générative ?	En utilisant le traitement par lots, le streaming de données, les transformations de données et l'ingénierie des fonctionnalités.
Comment les résultats de la solution d'IA générative seront-ils consommés ou intégrés aux systèmes en aval ?	Par le biais de points de terminaison d'API, de files d'attente de messages, de mises à jour de bases de données, etc.
Quels modèles d'intégration pilotés par les événements peuvent être utilisés pour la solution d'IA générative ?	Les files d'attente de messages (comme Amazon SQS, Apache Kafka, RabbitMQ), les systèmes pub/sub, les webhooks, les plateformes de streaming d'événements.
Quelles approches d'intégration basées sur les API peuvent être utilisées pour connecter la solution d'IA générative à d'autres systèmes ?	RESTful APIs, GraphQL APIs, SOAP APIs (pour les anciens systèmes).
Quels composants de l'architecture de microservices peuvent être utilisés pour l'intégration d'une solution d'IA générative ?	Maillage de services pour la communication interservices, les passerelles d'API, l'orchestration de conteneurs (par exemple, Kubernetes).
Comment l'intégration hybride peut-elle être mise en œuvre pour la solution d'IA générative ?	En combinant des modèles pilotés par les événements pour les mises à jour en temps réel, le traitement par lots des données historiques et APIs pour l'intégration de systèmes externes.
Comment les résultats de la solution d'IA générative peuvent-ils être intégrés aux systèmes en aval ?	Par le biais de points de terminaison d'API, de files d'attente de messages, de mises à jour de bases de données, de webhooks et d'exportations de fichiers.

Question	Exemple de réponse
Quelles mesures de sécurité doivent être prises en compte pour intégrer la solution d'IA générative ?	Mécanismes d'authentification (tels que OAuth JWT), chiffrement (en transit et au repos), limitation du débit des API et listes de contrôle d'accès (ACLs).
Comment prévoyez-vous d'intégrer des frameworks open source tels que LlamaIndex ou LangChain dans votre pipeline de données existant et votre flux de travail d'IA générative ?	Nous prévoyons de l'utiliser LangChain pour créer des applications d'IA génératives complexes, en particulier pour ses capacités de gestion des agents et de la mémoire. Notre objectif est de faire en sorte que 60 % de nos projets d'IA générative LangChain soient utilisés au cours des 6 prochains mois.
Comment garantirez-vous la compatibilité entre les frameworks open source que vous avez choisis et votre infrastructure de données existante ?	Nous sommes en train de créer une équipe d'intégration dédiée pour garantir une compatibilité fluide. D'ici le troisième trimestre , notre objectif est de disposer d'un pipeline entièrement intégré permettant d'indexer et de récupérer efficacement les données au sein de notre structure de lac de données actuelle. LlamaIndex
Comment comptez-vous tirer parti des composants modulaires des frameworks, par exemple LangChain pour le prototypage rapide et l'expérimentation ?	Nous mettons en place un environnement sandbox dans lequel les développeurs peuvent rapidement prototyper en utilisant LangChain les composants.
Quelle est votre stratégie pour suivre les mises à jour et les nouvelles fonctionnalités de ces frameworks open source en évolution rapide ?	Nous avons chargé une équipe de surveiller les GitHub référentiels et les forums communautaires pour LangChain et LlamaIndex. Nous prévoyons d'évaluer et d'intégrer des mises à jour majeures tous les trimestres, en mettant l'accent sur l'amélioration des performances et les nouvelles fonctionnalités.

Test

Question	Exemple de réponse
Quelles sont les exigences en matière de tests (par exemple, tests unitaires, tests d'intégration, end-to-end tests) ?	Tests unitaires pour des composants individuels, tests d'intégration avec des systèmes externes, end-to-end tests pour des scénarios critiques, etc.
Comment garantissez-vous la qualité et la cohérence des données entre les différentes sources pour la formation à l'IA générative ?	Nous maintenons la qualité des données grâce à des outils automatisés de profilage des données, à des audits de données réguliers et à un catalogue de données centralisé. Nous avons mis en place des politiques de gouvernance des données afin de garantir la cohérence entre les sources et de maintenir le lignage des données.
Comment le modèle d'IA générative sera-t-il évalué et validé ?	En utilisant un ensemble de données résistant, une évaluation humaine, des tests A/B, etc.
Quels sont les critères d'évaluation des performances et de la précision du modèle d'IA générative ?	Précision, rappel, score F1, perplexité, évaluation humaine, etc.
Comment les cas extrêmes et les cas critiques seront-ils identifiés et traités ?	En utilisant une suite de tests complète, une évaluation humaine, des tests contradictoires, etc.
Comment testerez-vous les biais potentiels dans le modèle d'IA générative ?	En utilisant une analyse de parité démographique, des tests d'égalité des chances, des techniques de débiais contradictoires, des tests contrefactuels, etc.
Quels paramètres seront utilisés pour mesurer l'équité des résultats du modèle ?	Rapport d'impact disparate, chances égalisées, parité démographique, indicateurs d'équité individuels, etc.

Question	Exemple de réponse
Comment garantirez-vous une représentation diversifiée dans vos ensembles de données de test pour la détection des biais ?	En utilisant un échantillonnage stratifié par groupes démographiques, en collaboration avec des experts en diversité, en utilisant des données synthétiques pour combler les lacunes, etc.
Quel processus sera mis en œuvre pour le suivi continu de l'équité des modèles après le déploiement ?	Audits d'équité réguliers, systèmes automatisés de détection des biais, analyse des commentaires des utilisateurs, formation continue périodique avec des ensembles de données mis à jour, etc.
Comment aborderez-vous les biais intersectionnels dans le modèle d'IA générative ?	En utilisant une analyse d'équité intersectionnelle, des tests en sous-groupes, une collaboration avec des experts du domaine sur l'intersectionnalité, etc.
Comment testerez-vous les performances du modèle dans différents contextes linguistiques et culturels ?	En utilisant des ensembles de tests multilingues, une collaboration avec des experts culturels, des mesures d'équité localisées, des études comparatives interculturelles, etc.

Déploiement et automatisation

Question	Exemple de réponse
Quelles sont les exigences en matière de mise à l'échelle et d'équilibrage de charge ?	Routage intelligent des demandes ; système de mise à l'échelle automatique ; optimisation pour les démarrages à froid rapides en utilisant des techniques telles que la mise en cache des modèles, le chargement différé et les systèmes de stockage distribués ; conception du système pour gérer des modèles de trafic rapides et imprévisibles.

Question	Exemple de réponse
Quelles sont les exigences relatives à la mise à jour et au déploiement de nouvelles versions ?	Déploiements bleu/vert, versions de Canary, mises à jour continues, etc.
Quelles sont les exigences en matière de reprise après sinistre et de continuité des activités ?	Procédures de sauvegarde et de restauration, mécanismes de basculement, configurations de haute disponibilité, etc.
Quelles sont les exigences pour automatiser la formation, le déploiement et la gestion du modèle d'IA générative ?	Pipeline de formation automatisé, déploiement continu, mise à l'échelle automatique, etc.
Comment le modèle d'IA générative sera-t-il mis à jour et réentraîné au fur et à mesure que de nouvelles données seront disponibles ?	Par le biais d'une reconversion périodique, d'un apprentissage progressif, d'un apprentissage par transfert, etc.
Quelles sont les exigences relatives à l'automatisation de la surveillance et de la gestion ?	Alertes automatisées, mise à l'échelle automatique, autoréparation, etc.
Quel est votre environnement de déploiement préféré pour les charges de travail génératives liées à l'IA ?	Une approche hybride qui utilise AWS pour la formation des modèles et notre infrastructure sur site pour l'inférence afin de répondre aux exigences de résidence des données.
Y a-t-il des plateformes cloud spécifiques que vous préférez pour les déploiements d'IA générative ?	Services AWS, en particulier Amazon SageMaker AI pour le développement et le déploiement de modèles, et Amazon Bedrock pour les modèles de base.
Quelles technologies de conteneurisation envisagez-vous pour les charges de travail génératives liées à l'IA ?	Nous voulons standardiser les conteneurs Docker orchestrés avec Kubernetes afin de garantir la portabilité et l'évolutivité dans notre environnement hybride.
Avez-vous des outils préférés pour le CI/CD dans votre pipeline d'IA générative ?	GitLab pour le contrôle de version et les pipelines CI/CD, intégrés à Jenkins pour les tests et les déploiements automatisés.

Question	Exemple de réponse
Quels outils d'orchestration envisagez-vous pour gérer les flux de travail d'IA générative ?	Apache Airflow pour l'orchestration des flux de travail, en particulier pour le prétraitement des données et les pipelines d'entraînement des modèles.
Avez-vous des exigences spécifiques en matière d'infrastructure sur site pour prendre en charge les charges de travail génératives liées à l'IA ?	Nous investissons dans des serveurs accélérés par GPU et dans des réseaux haut débit pour prendre en charge les charges de travail d'inférence sur site.
Comment prévoyez-vous de gérer le versionnement et le déploiement des modèles dans différents environnements ?	Nous prévoyons de l'utiliser MLflow pour le suivi des modèles et le versionnement, et de l'intégrer à notre infrastructure Kubernetes pour un déploiement fluide dans tous les environnements.
Quels outils de surveillance et d'observabilité envisagez-vous pour les déploiements d'IA générative ?	Prometheus pour la collecte des métriques et Grafana pour la visualisation, avec des solutions de journalisation personnalisées supplémentaires pour la surveillance spécifique au modèle.
Comment abordez-vous le mouvement et la synchronisation des données dans un modèle de déploiement hybride ?	Nous utiliserons AWS DataSync un transfert de données efficace entre le stockage sur site et AWS des tâches de synchronisation automatisées planifiées en fonction de nos cycles de formation.
Quelles mesures de sécurité mettez-vous en œuvre pour les déploiements d'IA générative dans différents environnements ?	Nous utiliserons l'IAM pour les ressources cloud, intégrées à notre Active Directory sur site pour implémenter le end-to-end chiffrement et la segmentation du réseau afin de sécuriser les flux de données.

Stratégie en matière de données

Question	Exemple de réponse
Quels types de données spécifiques sont essentiels pour vos charges de travail génératives liées à l'IA, et quel pourcentage d'entre elles sont actuellement accessibles ?	Les journaux d'appels des clients et les données relatives aux avis sur les produits sont essentiels. Actuellement, 85 % de ces types de données sont accessibles pour nos projets d'IA générative.
Comment garantissez-vous et mesurez-vous la qualité de vos données ?	Nous avons mis en place des mesures de qualité des données, notamment l'exhaustivité, l'exactitude, la cohérence et l'actualité. Nous utilisons des outils automatisés pour évaluer régulièrement ces indicateurs et disposons d'une équipe dédiée au nettoyage et à l'enrichissement des données.
Quel pourcentage de vos données répond à vos normes de qualité en matière d'utilisation de l'IA générative ?	Actuellement, 78 % de nos données répondent à nos normes de qualité. Nous visons 95 % au cours des 12 prochains mois grâce à de meilleurs processus de nettoyage des données.
Comment comptez-vous renforcer la confiance de vos parties prenantes quant à l'utilisation des données dans l'IA générative ?	Nous mettons en place un comité d'éthique de l'IA, fournissant des explications claires sur les décisions relatives à l'IA et menant des audits trimestriels sur l'IA pour garantir la transparence et l'équité.
Dans quelle mesure votre documentation relative aux sources de données et au lignage est-elle complète ?	Nous maintenons un catalogue de données détaillé qui inclut les métadonnées de toutes nos sources de données, y compris l'origine, la fréquence des mises à jour et l'utilisation. Nous utilisons des outils de traçabilité des données pour suivre la manière dont les données circulent et se transforment dans nos systèmes.

Question	Exemple de réponse
Comment garantissez-vous la diversité de vos ensembles de données afin d'éviter les biais dans les modèles d'IA ?	Nous nous approvisionnons activement en données provenant de divers groupes démographiques et auditions régulièrement nos ensembles de données pour détecter tout biais représentatif. Nous utilisons également des techniques de génération de données synthétiques pour équilibrer les catégories sous-représentées.
Quel est votre taux de rafraîchissement des données pour les modèles d'IA générative critiques, et comment déterminez-vous cette fréquence ?	Les modèles critiques sont actualisés chaque semaine. Cette fréquence est déterminée par les indicateurs de performance des tests A/B, et nous visons une dégradation maximale de 2 % entre les actualisations.
Combien de versions d'ensembles de données critiques maintenez-vous et pendant combien de temps ?	Nous maintenons les cinq dernières versions de chaque ensemble de données critique, avec une période de conservation de 18 mois pour chaque version.
Combien d'équipes interfonctionnelles participent à vos initiatives d'IA générative et ont accès à vos données ?	Nous avons trois équipes interfonctionnelles. Chaque équipe comprend des scientifiques des données, des experts du domaine, des éthiciens et des analystes commerciaux.
Quelles politiques et pratiques de gouvernance des données avez-vous mises en place ?	Nous avons un comité interfonctionnel de gouvernance des données qui supervise nos politiques en matière de données. Nous avons mis en place des contrôles d'accès basés sur les rôles, des systèmes de classification des données et des audits réguliers pour garantir le respect de notre cadre de gouvernance.

Question	Exemple de réponse
Quelles mesures avez-vous mises en place pour garantir la confidentialité des données, obtenir le consentement approprié et préserver la confidentialité ?	Nous avons mis en place un cadre complet de confidentialité des données aligné sur le RGPD et le CCPA. Cela inclut l'obtention d'un consentement explicite pour l'utilisation des données, la mise en œuvre de techniques d'anonymisation des données et des évaluations régulières de l'impact sur la vie privée.
Quel pourcentage de vos ensembles de données de formation à l'IA a été audité pour détecter tout biais au cours du dernier trimestre ?	70 % de nos ensembles de données de formation sur l'IA ont été audités pour détecter tout biais au dernier trimestre. Nous mettons en œuvre des outils automatisés de détection des biais pour atteindre 100 % d'audits trimestriels.
Quelle est votre capacité actuelle de traitement des données, et dans quelle mesure prévoyez-vous avoir besoin pour les futures charges de travail d'IA générative ?	Notre capacité actuelle est de 10 TB/day. We project needing 30 TB/day en un an et nous développons notre infrastructure pour répondre à cette demande.
Quelle est votre stratégie pour trouver un équilibre entre la confidentialité des données et les besoins en données des modèles d'IA générative ?	Nous mettons en œuvre des techniques avancées d'anonymisation et de génération de données synthétiques. Notre objectif est d'augmenter nos données utilisables pour l'IA de 40 % tout en réduisant les risques de confidentialité de 60 % au cours de la prochaine année.
Quel est le pourcentage de vos ensembles de données d'apprentissage automatique (ML) étiquetés avec précision, et quel est votre taux de précision cible ?	Actuellement, 85 % de nos ensembles de données ML sont étiquetés avec précision. Nous visons un taux de précision de 95 % au cours du prochain trimestre en utilisant des techniques d'étiquetage humaines et automatisées.

Traduire les informations issues des évaluations en résultats exploitables

Cette section fournit un cadre pour analyser les réponses au questionnaire et utiliser ces informations pour façonner l'architecture cible et les autres résultats clés de l'initiative de modernisation de l'IA générative. Ce cadre comble le fossé entre la collecte de données et la mise en œuvre, et garantit que l'évaluation oriente et oriente directement votre stratégie de modernisation.

Définition de l'architecture cible :

- Utilisez les réponses au questionnaire pour éclairer le choix des services cloud et la conception des pipelines de données.
- Assurez-vous que la conception de l'architecture prend en charge l'évolutivité et l'interopérabilité, comme indiqué dans le guide.

Évaluation de l'état de préparation du client :

- Analysez les réponses au questionnaire liées à l'infrastructure, aux processus et à la culture organisationnelle actuels.
- Identifiez les lacunes et créez un plan pour y remédier. Priorisez les lacunes essentielles au succès du MVP.

Cas d'utilisation et objectifs ambitieux :

- Extrayez les problèmes commerciaux spécifiques des réponses au questionnaire afin de définir des objectifs clairs en matière de cas d'utilisation.
- Définissez des objectifs ambitieux qui correspondent à la vision à long terme de votre organisation en matière de modernisation de l'IA générative.

Estimation de l'effort :

- Utilisez les données du questionnaire pour estimer les ressources, le temps et le budget nécessaires au MVP et à la mise en œuvre complète.
- Créez une approche progressive qui commence par le MVP et décrivez les phases suivantes.

Besoins d'habilitation :

- Sur la base des réponses au questionnaire, identifiez les lacunes en matière de compétences et les besoins de formation.
- Développez un plan de formation qui répond à la fois aux besoins immédiats des MVP et à l'adoption à long terme de l'IA générative.

Plan de mise en œuvre :

- Créez une feuille de route complète qui commence par le MVP et décrit les étapes à suivre pour une modernisation complète de l'IA générative.
- Définissez des jalons et des livrables clairs pour chaque phase de la mise en œuvre.

Étapes pratiques :

- Matrice de priorisation : créez une matrice qui fait correspondre les réponses au questionnaire aux [six résultats](#) pour aider à hiérarchiser les fonctionnalités et les efforts.
- Approche itérative : Concevez le MVP pour qu'il soit la première itération d'une série de versions planifiées, chaque version s'appuyant sur l'architecture cible complète.
- Harmonisation des parties prenantes : utilisez les résultats du questionnaire pour aligner les parties prenantes sur le champ d'application du MVP et sur l'approche progressive visant à atteindre tous les résultats.
- Boucle de feedback continue : mettez en œuvre des mécanismes pour recueillir des commentaires après le déploiement du MVP, et utilisez les informations pour affiner les plans pour les phases suivantes.
- Mise en œuvre agile : Adoptez une méthodologie agile qui permet de traiter avec flexibilité tous les résultats au fil du temps, en commençant par les résultats les plus critiques du MVP.

Étapes suivantes

Une fois que vous avez terminé l'évaluation de la charge de travail générative de l'IA, procédez comme suit :

1. Fournir une architecture cible détaillée
 - Objectif : L'architecte de solutions crée une architecture cible complète qui correspond aux objectifs de l'organisation et aux résultats de l'évaluation.
 - Composants : Cette architecture inclut la conception de l'ingestion des données, des points d'intégration et de l'interopérabilité du système afin de garantir l'évolutivité, la fiabilité et l'optimisation des performances.
2. Expliquez dans quelle mesure les spécificités du cas d'utilisation sont Services AWS adaptées
 - Cartographie des services : identifiez et cartographiez les services spécifiques Services AWS qui correspondent le mieux aux cas d'utilisation identifiés.
 - Avantages : mettez en évidence la manière dont ces services répondent aux besoins spécifiques de l'entreprise, améliorent l'efficacité et offrent une évolutivité.
3. Fournir des solutions alternatives facultatives avec des avantages et des inconvénients
 - Solutions de rechange : Présentez des solutions alternatives qui pourraient également répondre aux exigences de l'organisation.
 - Analyse : Proposez une analyse détaillée des avantages et des inconvénients de chaque alternative en tenant compte de facteurs tels que le coût, la complexité et l'alignement sur les objectifs commerciaux.
4. Fournir une estimation détaillée du prix de Services AWS
 - Analyse des coûts : fournissez une estimation détaillée des coûts pour le projet proposé Services AWS, y compris les scénarios d'utilisation potentiels et les modèles de tarification.
 - Harmonisation du budget : assurez-vous que le coût correspond aux contraintes budgétaires de l'organisation et qu'il permet de bien comprendre les implications financières.
5. Obtenez des commentaires sur l'architecture proposée
 - Engagement des parties prenantes : Interagissez avec les parties prenantes pour présenter l'architecture proposée et recueillir des commentaires.
 - Amélioration itérative : utilisez les commentaires pour affiner et améliorer la solution, et confirmer qu'elle répond aux besoins et aux attentes de toutes les parties prenantes.

FAQ

Quel est l'objectif principal de l'évaluation de la charge de travail de l'IA générative ?

L'objectif principal de l'évaluation est d'évaluer l'état de préparation d'une organisation à la modernisation de ses charges de travail génératives en matière d'IA, d'identifier les cas d'utilisation et de développer une architecture de solution cible. Il vise à définir les exigences de modernisation, à déterminer le périmètre de mise en œuvre et à préparer une modernisation réussie de l'IA générative.

Qui devrait utiliser cette évaluation ?

Cette évaluation s'adresse aux architectes de solutions, aux architectes d'entreprise et aux architectes d'applications qui souhaitent évaluer les aspects techniques de la modernisation de l'IA générative. Il est également utile aux responsables de programme et aux responsables du personnel pour évaluer l'état de préparation global, l'allocation des ressources et les besoins en matière d'habilitation.

Quels sont les principaux éléments évalués dans le cadre de l'évaluation ?

L'évaluation couvre l'état de préparation global, les cas d'utilisation, l'architecture, le stockage, les réglementations et la conformité, l'intégration, les tests, l'automatisation du déploiement et la stratégie de données. Ces composants sont essentiels pour déterminer le niveau de préparation technique et organisationnel en vue de l'adoption de la modernisation de l'IA générative.

Comment l'évaluation aide-t-elle à définir l'architecture cible ?

L'évaluation fournit une approche structurée pour évaluer les systèmes actuels et identifier les améliorations. Il vous aide à sélectionner les technologies appropriées et à concevoir des architectures évolutives conformes aux objectifs commerciaux et aux exigences des cas d'utilisation.

Quels sont les avantages d'une évaluation générative de la charge de travail liée à l'IA ?

Les avantages incluent une efficacité accrue, une meilleure prise de décision, l'assurance de la conformité, la promotion de l'innovation et la préparation à l'évolutivité. L'évaluation établit une approche stratégique de la modernisation de l'IA générative et maximise les avantages potentiels tout en atténuant les risques.

Comment les organisations peuvent-elles garantir une mise en œuvre réussie à la suite de l'évaluation ?

Organisations devraient élaborer un plan de mise en œuvre clair comprenant des étapes définies, impliquer les parties prenantes dès le début et adopter une approche itérative. La mise en place d'un centre d'excellence (CoE) et l'accent mis sur le développement des talents sont également des meilleures pratiques recommandées.

À quels défis les organisations pourraient-elles être confrontées lors de l'évaluation ?

Les défis peuvent inclure la résistance au changement, les problèmes de qualité des données et les complexités liées à la conformité. Pour relever ces défis, il faut favoriser une culture de l'innovation, garantir la disponibilité des données et mettre en œuvre des mesures de sécurité robustes.

Comment l'évaluation répond-elle aux exigences réglementaires et de conformité ?

L'évaluation évalue les mesures de conformité actuelles et identifie les lacunes. Il garantit que les solutions cibles respectent les réglementations pertinentes et les lois sur la confidentialité des données, et intègrent les meilleures pratiques de sécurité pour protéger les informations sensibles.

Quel est le rôle de l'engagement des parties prenantes dans le processus d'évaluation ?

L'engagement des parties prenantes est essentiel pour obtenir l'adhésion, aligner les initiatives de modernisation sur les objectifs commerciaux et garantir une mise en œuvre réussie. Une implication

précoce et une communication claire des avantages sont essentielles pour surmonter les résistances et favoriser le soutien.

Comment les entreprises peuvent-elles mesurer le succès de leurs initiatives de modernisation de l'IA générative après l'évaluation ?

Le succès peut être mesuré à l'aide d'indicateurs de performance clés (KPIs) qui correspondent aux objectifs commerciaux. Le suivi et l'évaluation réguliers de ces indicateurs aident à orienter la prise de décision et à démontrer la valeur de la modernisation de l'IA générative aux parties prenantes.

En quoi l'approche d'évaluation diffère-t-elle pour les organisations de différentes tailles (petites, moyennes ou entreprises) ou de différents secteurs d'activité ?

Petites organisations :

- Peut avoir des ressources et une expertise limitées pour des évaluations complètes
- Susceptible de se concentrer sur des cas d'utilisation spécifiques à fort impact plutôt que sur une adoption à l'échelle de l'entreprise
- Peut s'appuyer davantage sur des outils et services tiers pour l'évaluation
- Le processus d'évaluation pourrait être moins formel et plus souple

Entreprises de taille moyenne :

- Disposent souvent d'équipes informatiques ou de données dédiées, mais elles peuvent manquer d'expertise spécialisée en IA
- Pourrait adopter une approche progressive, en commençant par des projets pilotes dans les principaux départements
- Nécessité d'équilibrer l'innovation avec les systèmes et processus existants
- L'évaluation implique probablement des équipes interfonctionnelles

Organisations d'entreprise :

- Disposez généralement d' AI/ML équipes dédiées et de ressources supplémentaires pour une évaluation complète
- Nécessité d'envisager des intégrations complexes avec les systèmes d'entreprise existants
- Il peut y avoir des exigences réglementaires spécifiques à l'industrie à prendre en compte
- L'évaluation implique souvent des processus de gouvernance formels

Ressources

- [IA générative activée AWS](#)
- [AWS propose de nouveaux guides sur l'intelligence artificielle, l'apprentissage automatique et l'IA générative pour planifier votre stratégie en matière d'IA](#) (article de AWS blog)
- [Bonnes pratiques pour créer des applications d'IA génératives AWS](#) (article de AWS blog)
- [Générateur d'applications d'IA générative activé AWS](#) (bibliothèque de AWS solutions)
- [Capacités d'IA générative](#) (architecture AWS de référence de sécurité)
- [AWS cadre des meilleures pratiques en matière d'IA générative](#) (guide de AWS Audit Manager l'utilisateur)
- [Choisir un service d'IA génératif](#) (guide de AWS décision)
- [Qu'est-ce qu'Amazon Bedrock ?](#) (Guide de l'utilisateur d'Amazon Bedrock)
- [Qu'est-ce qu'Amazon SageMaker AI ?](#) (Guide du développeur Amazon SageMaker AI)

Historique du document

Le tableau suivant décrit les modifications importantes apportées à ce guide. Pour être averti des mises à jour à venir, abonnez-vous à un [fil RSS](#).

Modification	Description	Date
Publication initiale	—	6 novembre 2024

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.