Utilisation d'Amazon Comprehend Medical LLMs et pour les soins de santé et les sciences de la vie

AWS Directives prescriptives

AWS Directives prescriptives: Utilisation d'Amazon Comprehend Medical LLMs et pour les soins de santé et les sciences de la vie

Table of Contents

Utilisation d'Amazon Comprehend Medical LLMs et pour les soins de santé et les sciences de la vie

Amazon Web Services (???contributeurs)

Décembre 2025 (historique du document)

Présentation de

Le volume toujours croissant de données médicales et la nécessité d'un traitement efficace et précis ont entraîné l'adoption du traitement du langage naturel (NLP) avec les technologies d'intelligence artificielle et d'apprentissage automatique (AI/ML). Les modèles de classification préentraînés et les grands modèles linguistiques (LLMs) sont devenus de puissants outils pour diverses tâches de PNL médicale, notamment la réponse à des questions cliniques, la synthèse de rapports et la génération d'informations. Cependant, le domaine des soins de santé et des sciences de la vie présente des défis uniques en raison de la complexité de la terminologie médicale, des connaissances spécifiques au domaine et des exigences réglementaires. L'utilisation efficace de classificateurs préentraînés ou LLMs dans ce domaine nécessite une approche bien conçue qui combine les points forts de ces modèles avec des ressources et des techniques spécifiques au domaine.

Les pratiques du secteur des soins de santé et des sciences de la vie reposent traditionnellement sur des systèmes basés sur des règles, un codage manuel et des processus d'évaluation par des experts. Ces systèmes et processus prennent du temps et sont sujets aux erreurs. L'intégration des technologies d'IA et de PNL, telles qu'Amazon Comprehend Medical et les modèles de base d'Amazon Bedrock, offre des solutions efficaces et évolutives pour le traitement des données médicales tout en améliorant la précision et la cohérence.

Ce guide explore l'utilisation d'Amazon Comprehend Medical LLMs et l'automatisation intelligente dans le secteur de la santé. Il décrit les meilleures pratiques, les défis et les approches pratiques pour rationaliser le codage médical, l'extraction des informations sur les patients et les processus de synthèse des dossiers. En utilisant les fonctionnalités d'Amazon Comprehend Medical, LLMs les établissements de santé peuvent atteindre de nouveaux niveaux d'efficacité opérationnelle, réduire les coûts et potentiellement améliorer les soins aux patients.

Le guide détaille les considérations uniques du domaine de la santé, telles que la compréhension de la terminologie médicale, l'utilisation d'un domaine spécifique LLMs et la prise en compte des limites

des AI/ML systèmes. Il fournit un parcours décisionnel complet aux responsables informatiques, aux architectes et aux responsables techniques du secteur de la santé afin d'évaluer l'état de préparation de l'organisation, d'évaluer les options de mise en œuvre Services AWS et d'utiliser les outils appropriés pour une automatisation réussie.

En suivant les directives et les meilleures pratiques décrites dans ce guide, les établissements de santé peuvent exploiter le pouvoir des AI/ML technologies tout en maîtrisant les complexités du domaine médical. Cette approche soutient le respect des directives éthiques et réglementaires et promeut l'utilisation responsable des systèmes d'IA dans les soins de santé. Il est conçu pour générer des informations précises et confidentielles.

Public visé

Ce guide est destiné aux acteurs technologiques, aux architectes, aux responsables techniques et aux décideurs qui souhaitent mettre en œuvre des solutions de traitement du langage naturel basées sur l'IA pour l'analyse et l'automatisation des données médicales.

Objectifs

Les organisations du secteur de la santé et des sciences de la vie peuvent atteindre plusieurs objectifs commerciaux en utilisant Amazon Comprehend Medical LLMs et. Ces résultats incluent généralement l'augmentation de l'efficacité opérationnelle, la réduction des coûts et l'amélioration des soins aux patients. Cette section décrit les principaux objectifs commerciaux et les avantages associés à la mise en œuvre des stratégies et des meilleures pratiques décrites dans ce guide.

Voici certains des objectifs que les organisations peuvent atteindre en mettant en œuvre les directives et les meilleures pratiques de ce guide :

- Réduction du temps de développement — L'objectif ultime de ce guide est de réduire le temps de développement en fonction des coûts, de diminuer la dette technique et d'atténuer les échecs potentiels des projets liés au POC. En comprenant les AI/ML services clés, tels qu'Amazon Comprehend Medical, ainsi que les avantages et les limites de l'utilisation du LLM pour les tâches de santé, les entreprises peuvent accélérer la mise sur le marché et accélérer la réalisation de leurs objectifs commerciaux.
- Extraire des informations pour automatiser les tâches de codage médical — Après les visites des patients, les spécialistes du codage et les prestataires peuvent extraire des informations de textes médicaux, tels que des notes subjectives, objectives, d'évaluation et de plan (SOAP). Cela

peut réduire les efforts de documentation manuelle et aider le prestataire à se concentrer sur les besoins du patient. En combinant les fonctionnalités de reconnaissance d'entités d'Amazon Comprehend Medical LLMs, les entreprises peuvent extraire des informations médicales pertinentes à partir des dossiers des patients, des notes cliniques et d'autres sources de données de santé. Cela permet de minimiser les erreurs humaines et de promouvoir des pratiques cohérentes.

- Résumez les dossiers des patients et la documentation clinique — La synthèse automatisée des antécédents des patients, des plans de traitement et des résultats médicaux peut faire gagner un temps précieux aux prestataires de soins de santé. LLMs peut aider à générer une documentation clinique complète et structurée. Vous pouvez obtenir un contexte supplémentaire avec Amazon Comprehend Medical, utiliser un LLM dans un domaine médical ou peaufiner un LLM avec des données médicales. Ces approches peuvent aider à fournir des résumés précis et à garantir que la documentation est conforme aux exigences et aux normes de conformité.
- Soutenir les décisions cliniques et les soins aux patients : en utilisant des liens ontologiques dans Amazon Comprehend Medical LLMs, les prestataires peuvent répondre à des questions médicales ou demander des recommandations concernant les soins aux patients. Cela permet aux professionnels de santé de prendre des décisions éclairées qui améliorent les résultats pour les patients et réduisent le risque d'erreurs médicales.

Approches génératives de l'IA et de la PNL pour les soins de santé et les sciences de la vie

Le traitement du langage naturel (NLP) est une technologie d'apprentissage automatique qui permet aux ordinateurs d'interpréter, de manipuler et de comprendre le langage humain. Les organisations du secteur de la santé et des sciences de la vie disposent d'importants volumes de données provenant des dossiers des patients. Ils peuvent utiliser un logiciel NLP pour traiter automatiquement ces données. Par exemple, ils peuvent associer la PNL à l'IA générative pour rationaliser le codage médical, extraire des informations sur les patients et résumer les dossiers.

Selon la tâche NLP que vous souhaitez effectuer, différentes architectures peuvent être les mieux adaptées à votre cas d'utilisation. Ce guide aborde les options d'IA générative et de PNL suivantes pour les applications des soins de santé et des sciences de la vie sur AWS :

- Utilisation d'Amazon Comprehend Medical— Découvrez comment utiliser Amazon Comprehend Medical de manière indépendante, sans l'intégrer à un grand modèle de langage (LLM).
- Combiner Amazon Comprehend Medical avec de grands modèles linguistiques— Découvrez comment associer Amazon Comprehend Medical à un LLM dans une architecture RAG (Retrieval Augment Generation).
- Utilisation de grands modèles linguistiques pour les cas d'utilisation dans le domaine de la santé et des sciences de la vie— Découvrez comment utiliser un LLM pour les applications de santé et des sciences de la vie, soit en utilisant un LLM affiné, soit une architecture RAG.

Utilisation d'Amazon Comprehend Medical

Amazon Comprehend Medical détecte et renvoie des informations utiles dans des textes cliniques non structurés, tels que des notes du médecin, des résumés de sortie, des résultats de tests et des notes de cas. Service AWS Il utilise des modèles de traitement du langage naturel (NLP) pour détecter les entités. Les entités sont des références textuelles à des informations médicales, telles que des problèmes de santé, des médicaments ou des informations de santé protégées (PHI).

> **Important**
>
> Amazon Comprehend Medical ne remplace pas un avis médical, un diagnostic ou un traitement professionnel. Amazon Comprehend Medical fournit des scores de confiance

qui indiquent le niveau de confiance dans la précision des entités détectées. Déterminez le seuil de confiance approprié pour votre cas et utilisez des seuils de confiance élevés dans les situations qui exigent une grande précision. Dans certains cas d'utilisation, les résultats doivent être examinés et vérifiés par des évaluateurs humains dûment formés. Par exemple, Amazon Comprehend Medical ne doit être utilisé dans des scénarios de soins aux patients qu'après vérification de l'exactitude et du bon jugement médical par des professionnels de santé qualifiés.

Vous pouvez accéder à Amazon Comprehend Medical via AWS Management Console le, AWS Command Line Interface le AWS CLI() ou via AWS SDKs le. Ils AWS SDKs sont disponibles pour différents langages de programmation et plateformes, tels que Java, Python, Ruby, .NET, iOS et Android. Vous pouvez utiliser le SDKs pour accéder par programmation à Amazon Comprehend Medical depuis votre application client.

Cette section passe en revue les principales fonctionnalités d'Amazon Comprehend Medical. Il décrit également les avantages de l'utilisation de ce service par rapport à un grand modèle linguistique (LLM).

Fonctionnalités d'Amazon Comprehend Medical

Amazon Comprehend Medical APIs propose une inférence en temps quasi réel et par lots. Ils APIs peuvent ingérer du texte médical et fournir des résultats pour les tâches de PNL médicale en utilisant la reconnaissance des entités médicales et en identifiant les relations entre les entités. Vous pouvez effectuer une analyse à la fois sur des fichiers individuels ou sous forme d'analyse par lots sur plusieurs fichiers stockés dans un bucket Amazon Simple Storage Service (Amazon S3). Amazon Comprehend Medical propose les opérations d'API d'analyse de texte suivantes pour la détection synchrone des entités :

- Détecter les entités : détecte les catégories médicales générales telles que l'anatomie, l'état de santé, la catégorie PHI, les procédures et les expressions temporelles.
- Détecter les PHI — Détecte des entités spécifiques telles que l'âge, la date, le nom et des informations personnelles similaires.

Amazon Comprehend Medical inclut également plusieurs opérations d'API que vous pouvez utiliser pour effectuer une analyse de texte par lots sur des documents cliniques. Pour en savoir plus sur l'utilisation de ces opérations d'API, consultez la section Analyse de texte par lots APIs.

Utilisez Amazon Comprehend Medical pour détecter des entités dans un texte clinique et relier ces entités à des concepts issus d'ontologies médicales standardisées, notamment RxNorm les bases de connaissances ICD-10-CM et SNOMED CT. Vous pouvez effectuer une analyse à la fois sur des fichiers individuels ou sous forme d'analyse par lots sur des documents volumineux ou sur plusieurs fichiers stockés dans un compartiment Amazon S3. Amazon Comprehend Medical propose l'ontologie suivante reliant les opérations d'API :

- Infer ICD10 CM — L'opération Infer ICD10 CM détecte les affections médicales potentielles et les relie aux codes de la version 2019 de la Classification internationale des maladies, 10e révision, modification clinique (ICD-10-CM). Pour chaque problème médical potentiel détecté, Amazon Comprehend Medical répertorie les codes et descriptions ICD-10-CM correspondants. Les affections médicales répertoriées dans les résultats incluent un score de confiance, qui indique la confiance d'Amazon Comprehend Medical dans la précision des entités par rapport aux concepts correspondants dans les résultats.
- InferRxNorm— L'InferRxNormopération identifie les médicaments répertoriés dans le dossier d'un patient en tant qu'entités. Il relie les entités aux identificateurs de concepts (RxCUI) de la RxNorm base de données de la National Library of Medicine. Chaque RxCUI est unique pour différents dosages et formes posologiques. Les médicaments listés dans les résultats incluent un score de confiance, qui indique la confiance d'Amazon Comprehend Medical dans la précision des entités correspondant aux concepts de RxNorm la base de connaissances. Amazon Comprehend Medical répertorie les meilleurs CUIs Rx susceptibles de correspondre à chaque médicament détecté par ordre décroissant en fonction du score de confiance.
- InfersNoMedCT — L'opération InfersNoMedCT identifie les concepts médicaux possibles en tant qu'entités et les relie aux codes de la version 2021-03 de la Nomenclature systématisée de la médecine, termes cliniques (SNOMED CT). SNOMED CT fournit un vocabulaire complet de concepts médicaux, y compris les affections médicales et l'anatomie, ainsi que les tests médicaux, les traitements et les procédures. Pour chaque identifiant de concept correspondant, Amazon Comprehend Medical renvoie les cinq principaux concepts médicaux, chacun avec un score de confiance et des informations contextuelles telles que les traits et les attributs. Le concept SNOMED CT IDs peut ensuite être utilisé pour structurer les données cliniques des patients à des fins de codage médical, de reporting ou d'analyse clinique lorsqu'il est utilisé avec la polyhiérarchie SNOMED CT.

Pour plus d'informations, consultez Analyse de texte APIs et association d'ontologies APIs dans la documentation Amazon Comprehend Medical.

Cas d'utilisation d'Amazon Comprehend Medical

En tant que service autonome, Amazon Comprehend Medical peut répondre au cas d'utilisation de votre entreprise. Amazon Comprehend Medical peut effectuer des tâches telles que les suivantes :

- Aide au codage médical dans les dossiers des patients
- Détecter les données de santé protégées (PHI)
- Validation des médicaments, y compris les attributs tels que la posologie, la fréquence et la forme

Les résultats d'Amazon Comprehend Medical sont assimilables pour la majorité des cabinets médicaux. Toutefois, vous devrez peut-être envisager d'autres solutions si vous avez des limitations telles que les suivantes :

- Différentes définitions d'entités — Par exemple, votre définition `FREQUENCY` d'une entité médicamenteuse peut être différente. Pour ce qui est de la fréquence, Amazon Comprehend Medical prévoit en fonction des besoins, mais votre organisation peut utiliser le terme pro re nata (PRN).
- Une quantité impressionnante de résultats — Par exemple, les notes des patients contiennent souvent plusieurs symptômes et des mots clés correspondant à plusieurs codes ICD-10-CM. Cependant, plusieurs mots clés ne sont pas applicables au diagnostic. Dans ce cas, le fournisseur doit évaluer de nombreuses entités ICD-10-CM et leurs scores de confiance, ce qui nécessite un temps de traitement manuel.
- Entités personnalisées ou tâches NLP : par exemple, les prestataires peuvent vouloir extraire des preuves du PRN, par exemple en cas de douleur. Comme ce produit n'est pas disponible via Amazon Comprehend Medical, un AI/ML autre modèle est justifié. Une AI/ML solution différente est requise si la tâche NLP ne relève pas de la reconnaissance des entités, telle que la synthèse, la réponse aux questions et l'analyse des sentiments.

Combiner Amazon Comprehend Medical avec de grands modèles linguistiques

Une étude réalisée en 2024 par NEJM AI a montré que l'utilisation d'un LLM, avec indication « zero shot », pour les tâches de codage médical entraîne généralement de mauvaises performances. L'utilisation d'Amazon Comprehend Medical avec un LLM peut contribuer à atténuer ces problèmes de performances. Les résultats d'Amazon Comprehend Medical constituent un contexte utile pour

un LLM qui effectue des tâches de PNL. Par exemple, le fait de fournir un contexte allant d'Amazon Comprehend Medical au modèle linguistique étendu peut vous aider à :

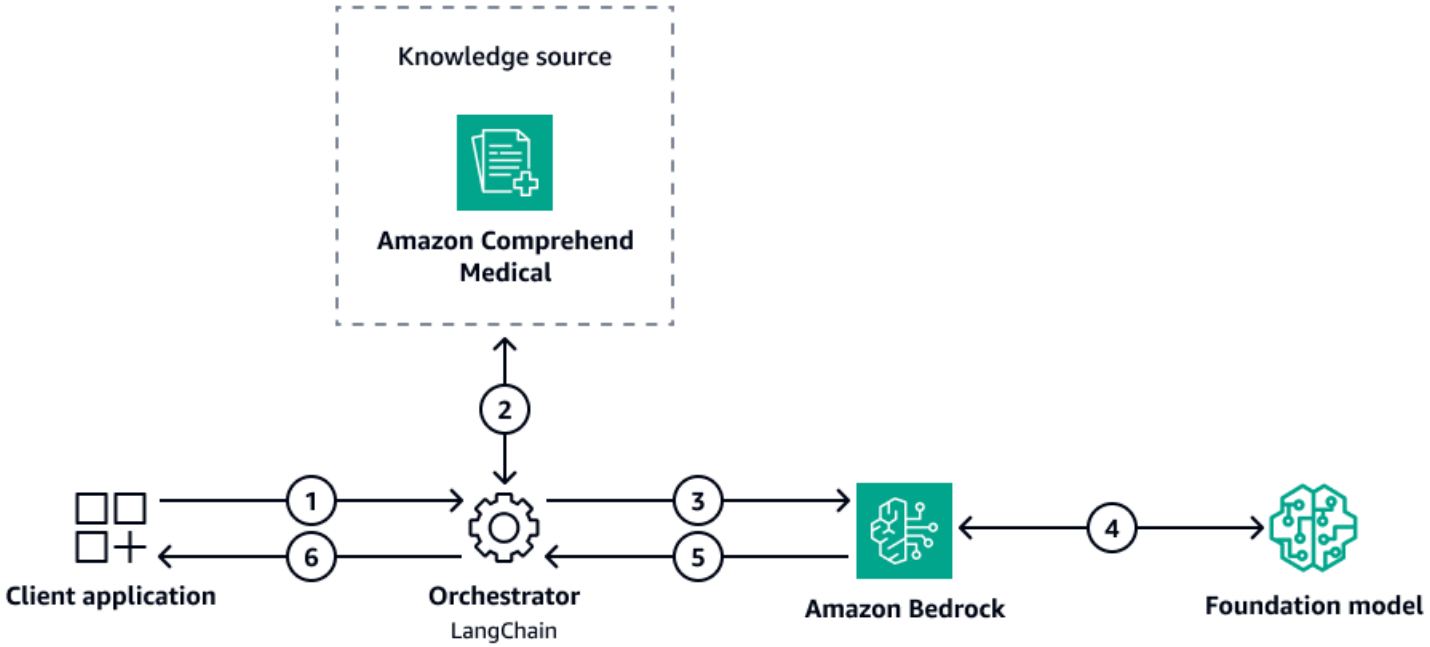
- Améliorez la précision des sélections d'entités en utilisant les premiers résultats d'Amazon Comprehend Medical comme contexte pour le LLM
- Implémentez la reconnaissance d'entités personnalisées, la synthèse, la réponse aux questions et des cas d'utilisation supplémentaires

Cette section explique comment associer Amazon Comprehend Medical à un LLM en utilisant une approche RAG (Retrieval Augmented Generation). La génération augmentée de récupération (RAG) est une technologie d'IA générative dans laquelle un LLM fait référence à une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation avant de générer une réponse. Pour plus d'informations, voir Qu'est-ce que RAG ?

Pour illustrer cette approche, cette section utilise l'exemple du codage médical (diagnostique) lié à la CIM-10-CM. Il inclut un exemple d'architecture et des modèles d'ingénierie rapides pour vous aider à accélérer votre innovation. Il inclut également les meilleures pratiques pour utiliser Amazon Comprehend Medical dans un flux de travail RAG.

Architecture basée sur RAG avec Amazon Comprehend Medical

Le schéma suivant illustre une approche RAG pour identifier les codes de diagnostic ICD-10-CM à partir des notes des patients. Il utilise Amazon Comprehend Medical comme source de connaissances. Dans une approche RAG, la méthode de récupération extrait généralement des informations d'une base de données vectorielle contenant les connaissances applicables. Au lieu d'une base de données vectorielle, cette architecture utilise Amazon Comprehend Medical pour la tâche de récupération. L'orchestrateur envoie les informations de la note du patient à Amazon Comprehend Medical et récupère les informations du code ICD-10-CM. L'orchestrateur envoie ce contexte au modèle de base en aval (LLM), via Amazon Bedrock. Le LLM génère une réponse en utilisant les informations du code ICD-10-CM, et cette réponse est renvoyée à l'application cliente.



Le diagramme montre le flux de travail RAG suivant :

1. L'application cliente envoie les notes du patient sous forme de requête à l'orchestrateur. Voici un exemple de ces remarques à l'intention des patients : « La patiente est une patiente de 71 ans du Dr X. La patiente s'est présentée aux urgences hier soir avec des douleurs abdominales persistantes depuis environ 7 à 8 jours. Elle n'a pas eu de fièvres ou de frissons précis et n'a aucun antécédent de jaunisse. Le patient nie toute perte de poids récente significative. »
2. L'orchestrateur utilise Amazon Comprehend Medical pour récupérer les codes ICD-10-CM relatifs aux informations médicales contenues dans la requête. Il utilise l'API Infer ICD10 CM pour extraire et déduire les codes ICD-10-CM à partir des notes du patient.
3. L'orchestrateur crée une invite qui inclut le modèle d'invite, la requête d'origine et les codes ICD-10-CM extraits d'Amazon Comprehend Medical. Il envoie ce contexte amélioré à Amazon Bedrock.
4. Amazon Bedrock traite les entrées et utilise un modèle de base pour générer une réponse qui inclut les codes ICD-10-CM et les preuves correspondantes issues de la requête. La réponse générée inclut les codes ICD-10-CM identifiés et les preuves provenant des notes du patient qui soutiennent chaque code. Voici un exemple de réponse :

```
<response>
<icd10>
<code>R10.9</code>
<evidence>history of abdominal pain</evidence>
```

```
</icd10>
<icd10>
<code>R10.30</code>
<evidence>history of abdominal pain</evidence>
</icd10>
</response>
```

5. Amazon Bedrock envoie la réponse générée à l'orchestrateur.
6. L'orchestrateur renvoie la réponse à l'application cliente, où l'utilisateur peut consulter la réponse.

Cas d'utilisation d'Amazon Comprehend Medical dans un flux de travail RAG

Amazon Comprehend Medical peut effectuer des tâches de PNL spécifiques. Pour plus d'informations, consultez la section Cas d'utilisation d'Amazon Comprehend Medical.

Vous souhaiterez peut-être intégrer Amazon Comprehend Medical dans un flux de travail RAG pour les cas d'utilisation avancés, tels que les suivants :

- Générez des résumés cliniques détaillés en combinant des entités médicales extraites avec des informations contextuelles issues des dossiers des patients
- Automatisez le codage médical pour les cas complexes en utilisant des entités extraites avec des informations liées à l'ontologie pour l'attribution du code
- Automatisez la création de notes cliniques structurées à partir de texte non structuré en utilisant des entités médicales extraites
- Analyser les effets secondaires des médicaments en fonction des noms et des attributs des médicaments extraits
- Développez des systèmes de soutien clinique intelligents qui combinent les informations médicales extraites avec up-to-date la recherche et les directives

Bonnes pratiques pour utiliser Amazon Comprehend Medical dans un flux de travail RAG

Lorsque vous intégrez les résultats d'Amazon Comprehend Medical dans une demande de LLM, il est essentiel de suivre les meilleures pratiques. Cela peut améliorer les performances et la précision. Les principales recommandations sont les suivantes :

- Comprendre les scores de confiance d'Amazon Comprehend Medical — Amazon Comprehend Medical fournit des scores de confiance pour chaque entité détectée et pour chaque lien d'ontologie. Il est essentiel de comprendre la signification de ces scores et d'établir des seuils appropriés pour votre cas d'utilisation spécifique. Les scores de confiance aident à filtrer les entités peu fiables, à réduire le bruit et à améliorer la qualité des entrées du LLM.
- Utilisez les scores de confiance pour une ingénierie rapide : lors de l'élaboration des instructions pour le LLM, pensez à intégrer les scores de confiance d'Amazon Comprehend Medical comme contexte supplémentaire. Cela permet au LLM de hiérarchiser ou d'évaluer les entités en fonction de leur niveau de confiance, améliorant ainsi potentiellement la qualité du résultat.
- Évaluez les résultats d'Amazon Comprehend Medical à l'aide de données fiables sur le terrain — Les données fiables sur le terrain sont des informations dont la véracité est reconnue. Il peut être utilisé pour valider qu'une AI/ML application produit des résultats précis. Avant d'intégrer les résultats d'Amazon Comprehend Medical dans votre flux de travail LLM, évaluez les performances du service sur un échantillon représentatif de vos données. Comparez les résultats avec des annotations fondées sur la vérité pour identifier les divergences potentielles ou les domaines à améliorer. Cette évaluation vous aide à comprendre les points forts et les limites d'Amazon Comprehend Medical pour votre cas d'utilisation.
- Sélectionnez stratégiquement les informations pertinentes : Amazon Comprehend Medical peut fournir une grande quantité d'informations, mais elles ne sont peut-être pas toutes pertinentes pour votre tâche. Sélectionnez avec soin les entités, les attributs et les métadonnées les plus pertinents pour votre cas d'utilisation. Fournir trop d'informations non pertinentes au LLM peut introduire du bruit et potentiellement réduire les performances.
- Aligner les définitions des entités : assurez-vous que les définitions des entités et des attributs utilisés par Amazon Comprehend Medical correspondent à votre interprétation. En cas de divergence, pensez à fournir un contexte ou des éclaircissements supplémentaires au LLM afin de combler le fossé entre les résultats d'Amazon Comprehend Medical et vos besoins. Si l'entité Amazon Comprehend Medical ne répond pas à vos attentes, vous pouvez implémenter une détection d'entité personnalisée en incluant des instructions supplémentaires (et des exemples possibles) dans l'invite.
- Fournissez des connaissances spécifiques à un domaine — Amazon Comprehend Medical fournit des informations médicales précieuses, mais il se peut qu'il ne capture pas toutes les nuances de votre domaine spécifique. Envisagez de compléter les résultats d'Amazon Comprehend Medical par des sources de connaissances supplémentaires spécifiques au domaine, telles que des ontologies, des terminologies ou des ensembles de données sélectionnés par des experts. Cela fournit un contexte plus complet au LLM.

- Respectez les directives éthiques et réglementaires — Lorsque vous traitez des données médicales, il est important de respecter les principes éthiques et les directives réglementaires, tels que ceux liés à la confidentialité des données, à la sécurité et à l'utilisation responsable des systèmes d'IA dans les soins de santé. Assurez-vous que votre mise en œuvre est conforme aux lois applicables et aux meilleures pratiques du secteur.

En suivant ces meilleures pratiques, les AI/ML praticiens peuvent utiliser efficacement les points forts d'Amazon Comprehend Medical LLMs et de. Pour les tâches de PNL médicale, ces meilleures pratiques permettent d'atténuer les risques potentiels et d'améliorer les performances.

Ingénierie rapide pour le contexte Amazon Comprehend Medical

L'ingénierie rapide est le processus qui consiste à concevoir et à affiner des instructions pour guider une solution d'IA générative afin de générer les résultats souhaités. Vous choisissez les formats, les phrases, les mots et les symboles les plus appropriés pour aider l'IA à interagir avec vos utilisateurs de manière plus significative.

En fonction de l'opération d'API que vous effectuez, Amazon Comprehend Medical renvoie les entités détectées, les codes et descriptions d'ontologie, ainsi que les scores de confiance. Ces résultats deviennent contextuels dans l'invite lorsque votre solution invoque le LLM cible. Vous devez concevoir l'invite de manière à présenter le contexte dans le modèle d'invite.

> **Note**
>
> Les exemples d'instructions de cette section suivent les directives d'Anthropic. Si vous utilisez un autre fournisseur de LLM, suivez les recommandations de ce fournisseur.

En général, vous insérez à la fois le texte médical d'origine et les résultats d'Amazon Comprehend Medical dans l'invite. Voici une structure d'invite courante :

```
<medical_text>
medical text
</medical_text>

<comprehend_medical_text_results>
comprehend medical text results
</comprehend_medical_text_results>
```

```
<prompt_instructions>
prompt instructions
</prompt_instructions>
```

Cette section fournit des stratégies pour inclure les résultats d'Amazon Comprehend Medical comme contexte rapide pour les tâches de PNL médicale courantes suivantes :

- Filtrer les résultats d'Amazon Comprehend Medical
- Étendez les tâches de PNL médicale avec Amazon Comprehend Medical
- Appliquez des garde-corps avec Amazon Comprehend Medical

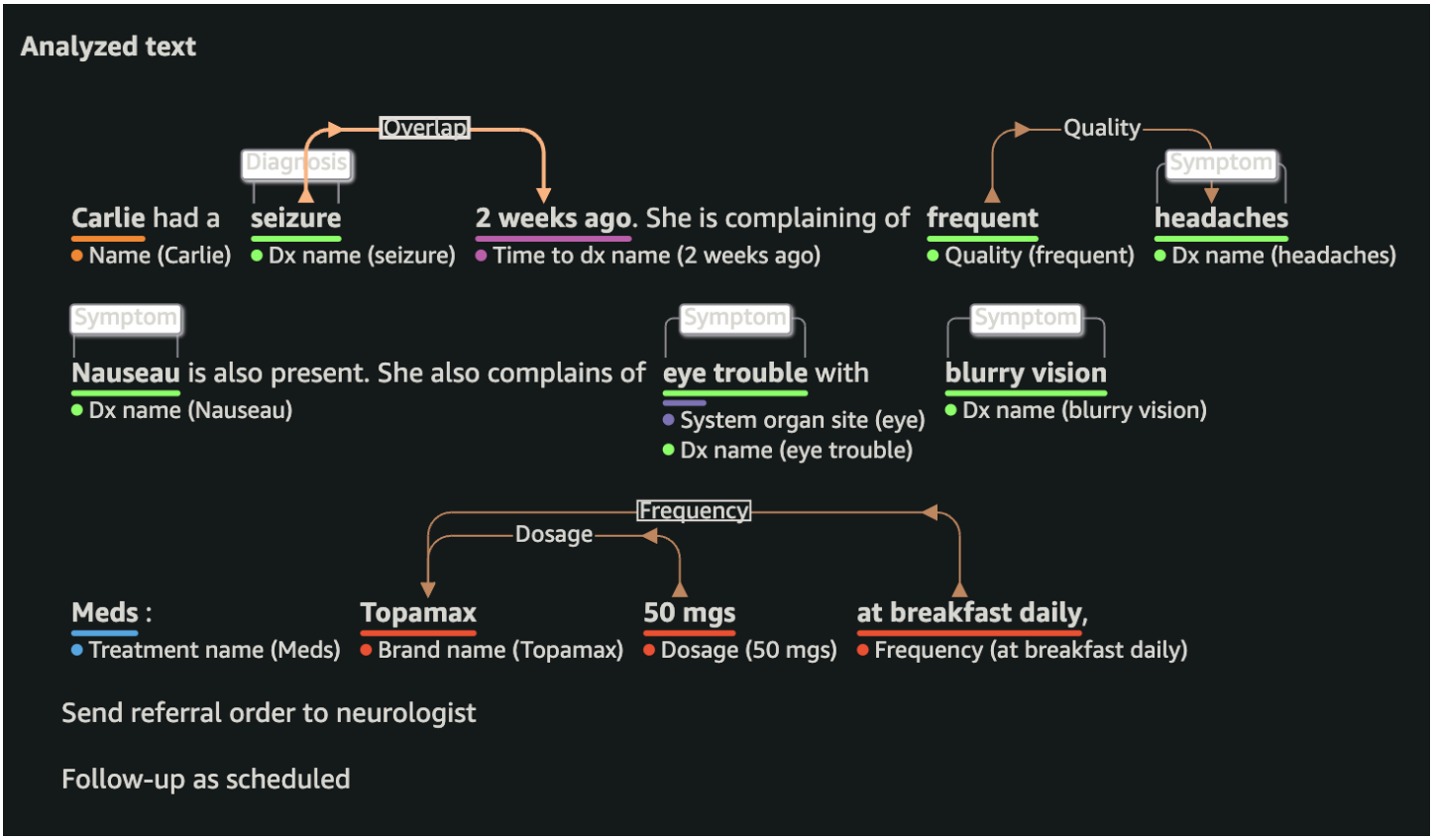
Filtrer les résultats d'Amazon Comprehend Medical

Amazon Comprehend Medical fournit généralement une grande quantité d'informations. Vous souhaiterez peut-être réduire le nombre de résultats que le professionnel de santé doit examiner. Dans ce cas, vous pouvez utiliser un LLM pour filtrer ces résultats. Les entités Amazon Comprehend Medical incluent un score de confiance que vous pouvez utiliser comme mécanisme de filtrage lors de la conception de l'invite.

Voici un exemple de note destinée aux patients :

```
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
```

Dans cette note destinée au patient, Amazon Comprehend Medical détecte les entités suivantes.



Les entités renvoient aux codes ICD-10-CM suivants pour les crises d'épilepsie et les maux de tête.

Catégorie	Code ICD-10-CM	Description de l'ICD-10-CM	Score de fiabilité
Convulsion	R56,9	Convulsions non précisées	0,8348
Convulsion	G40,909	Épilepsie, non précisée, non incurable, sans état épileptique	0,5424
Convulsion	R56,00	Convulsions fébriles simples	0,4937
Convulsion	G40,09	Autres crises	0,4397

Convulsion	G40,409	Autres syndromes épileptiques et épileptiques généralisés, non incurables, sans état épileptique	0,4138
maux de tête	R51	maux de tête	0,4067
maux de tête	R51,9	Céphalée, sans précision	0,3844
maux de tête	G44,52	Nouveaux maux de tête persistants quotidiens (NDPH)	0,3005
maux de tête	G44	Autre syndrome de céphalée	0,2670
maux de tête	G44,8	Autres syndromes de céphalée précisés	0,2542

Vous pouvez transmettre les codes ICD-10-CM à l'invite pour augmenter la précision du LLM. Pour réduire le bruit, vous pouvez filtrer les codes ICD-10-CM en utilisant le score de confiance inclus dans les résultats d'Amazon Comprehend Medical. Voici un exemple d'invite qui inclut uniquement les codes ICD-10-CM dont le score de confiance est supérieur à 0,4 :

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
</patient_note>

<comprehend_medical_results>
<icd-10>
  <entity>
    <text>seizure</text>
    <code>
```

```
      <description>Unspecified convulsions</description>
      <code_value>R56.9</code_value>
      <score>0.8347607851028442</score>
    </code>
    <code>
      <description>Epilepsy, unspecified, not intractable, without status epilepticus</
description>
      <code_value>G40.909</code_value>
      <score>0.542376697063446</score>
    </code>
    <code>
      <description>Other seizures</description>
      <code_value>G40.89</code_value>
      <score>0.43966275453567505</score>
    </code>
    <code>
      <description>Other generalized epilepsy and epileptic syndromes, not intractable,
 without status epilepticus</description>
      <code_value>G40.409</code_value>
      <score>0.41382506489753723</score>
    </code>
  </entity>
  <entity>
    <text>headaches</text>
    <code>
      <description>Headache</description>
      <code_value>R51</code_value>
      <score>0.4066613018512726</score>
    </code>
  </entity>
  <entity>
    <text>Nausea</text>
    <code>
      <description>Nausea</description>
      <code_value>R11.0</code_value>
      <score>0.6460834741592407</score>
    </code>
  </entity>
  <entity>
    <text>eye trouble</text>
    <code>
      <description>Unspecified disorder of eye and adnexa</description>
      <code_value>H57.9</code_value>
```

```
      <score>0.6780954599380493</score>
    </code>
    <code>
      <description>Unspecified visual disturbance</description>
      <code_value>H53.9</code_value>
      <score>0.5871203541755676</score>
    </code>
    <code>
      <description>Unspecified disorder of binocular vision</description>
      <code_value>H53.30</code_value>
      <score>0.5539672374725342</score>
    </code>
  </entity>
  <entity>
    <text>blurry vision</text>
    <code>
      <description>Other visual disturbances</description>
      <code_value>H53.8</code_value>
      <score>0.9001834392547607</score>
    </code>
  </entity>
</icd-10>
</comprehend_medical_results>

<prompt>
Given the patient note and Amazon Comprehend Medical ICD-10-CM code results above,
 please select the most relevant ICD-10-CM diagnosis codes for the patient.
For each selected code, provide a brief explanation of why it is relevant based on the
 information in the patient note.
</prompt>
```

Étendez les tâches de PNL médicale avec Amazon Comprehend Medical

Lors du traitement de textes médicaux, le contexte d'Amazon Comprehend Medical peut aider le LLM à sélectionner de meilleurs jetons. Dans cet exemple, vous souhaitez associer les symptômes diagnostiques aux médicaments. Vous souhaitez également trouver du texte relatif à des tests médicaux, tels que des termes relatifs à un test sanguin. Vous pouvez utiliser Amazon Comprehend Medical pour détecter les entités et les noms des médicaments. Dans ce cas, vous utiliserez le DetectEntitiesV2 et InferRxNorm APIs pour Amazon Comprehend Medical.

Voici un exemple de note destinée aux patients :

```
Carlie had a seizure 2 weeks ago. She is complaining of increased frequent headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
```

Pour vous concentrer sur le code de diagnostic, seules les entités associées au type `MEDICAL_CONDITION` with `DX_NAME` sont utilisées dans l'invite. Les autres métadonnées sont exclues en raison de leur non-pertinence. Pour les entités médicamenteuses, le nom du médicament ainsi que les attributs extraits sont inclus. Les autres métadonnées d'entités médicamenteuses d'Amazon Comprehend Medical sont exclues en raison de leur non-pertinence. Voici un exemple d'invite qui utilise les résultats filtrés d'Amazon Comprehend Medical. L'invite se concentre sur `MEDICAL_CONDITION` les entités qui possèdent ce `DX_NAME` type. Cette invite est conçue pour relier plus précisément les codes de diagnostic aux médicaments et pour extraire plus précisément les tests d'ordonnance médicale :

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of increased freqeunt headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order
 prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
</patient_note>

<detect_entity_results>
<entity>
    <text>seizure</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>headaches</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>lyme disease</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>muscle ache</text>
```

```
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>stiff neck</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
</detect_entity_results>

<rx_results>
<entity>
    <text>Topamax</text>
    <category>MEDICATION</category>
    <type>BRAND_NAME</type>
    <attributes>
        <attribute>
            <type>FREQUENCY</type>
            <text>at breakfast daily</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>50 mgs</text>
        </attribute>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
    </attributes>
</entity>
<entity>
    <text>Amoxicillan</text>
    <category>MEDICATION</category>
    <type>GENERIC_NAME</type>
    <attributes>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>25 mg</text>
        </attribute>
```

```
        <attribute>
            <type>FREQUENCY</type>
            <text>twice a day</text>
        </attribute>
    </attributes>
</entity>
</rx_results>

<prompt>
Based on the patient note and the detected entities, can you please:
1. Link the diagnosis symptoms with the medications prescribed.
Provide your reasoning for the linkages.
2. Extract any entities related to medical order tests mentioned in the note.
</prompt>
```

Appliquez des garde-corps avec Amazon Comprehend Medical

Vous pouvez utiliser un LLM et Amazon Comprehend Medical pour créer des barrières de sécurité avant que la réponse générée ne soit utilisée. Vous pouvez exécuter ce flux de travail sur du texte médical non modifié ou post-traité. Les cas d'utilisation incluent le traitement des informations de santé protégées (PHI), la détection d'hallucinations ou la mise en œuvre de politiques personnalisées pour la publication des résultats. Par exemple, vous pouvez utiliser le contexte d'Amazon Comprehend Medical pour identifier les données PHI, puis utiliser le LLM pour supprimer ces données PHI.

Voici un exemple d'informations extraites du dossier d'un patient contenant des informations médicales médicales :

```
Patient name: John Doe
Patient SSN: 123-34-5678
Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
```

Voici un exemple d'invite qui inclut les résultats d'Amazon Comprehend Medical comme contexte :

```
<original_text>
Patient name: John Doe
Patient SSN: 123-34-5678 Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
```

```
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
</original_text>

<comprehend_medical_phi_entities>
<entity>
  <text>John Doe</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9967944025993347</score>
  <type>NAME</type>
</entity>
<entity>
  <text>123-34-5678</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9998034834861755</score>
  <type>ID</type>
</entity>
<entity>
  <text>01/01/2000</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9964448809623718</score>
  <type>DATE</type>
</entity>
</comprehend_medical_phi_entities>

<instructions>
Using the provided original text and the Amazon Comprehend Medical PHI entities
 detected, please analyze the text to determine if it contains any additional protected
 health information (PHI) beyond the entities already identified. If additional PHI is
 found, please list and categorize it. If no additional PHI is found, please state that
 explicitly.
In addition if PHI is found, generate updated text with the PHI removed.
</instructions>
```

Utilisation de grands modèles linguistiques pour les cas d'utilisation dans le domaine de la santé et des sciences de la vie

Cela décrit comment vous pouvez utiliser de grands modèles linguistiques (LLMs) pour les applications de santé et des sciences de la vie. Certains cas d'utilisation nécessitent l'utilisation d'un modèle de langage étendu pour les fonctionnalités génératives de l'IA. Il y a des avantages et des limites, même pour le plus state-of-the-art LLMs grand nombre, et les recommandations de cette section sont conçues pour vous aider à atteindre les résultats que vous vous êtes fixés.

Vous pouvez utiliser le chemin de décision pour déterminer la solution LLM adaptée à votre cas d'utilisation, en tenant compte de facteurs tels que la connaissance du domaine et les données de formation disponibles. De plus, cette section traite des pratiques médicales préformées populaires LLMs et des meilleures pratiques pour leur sélection et leur utilisation. Il aborde également les compromis entre des solutions complexes et performantes et des approches plus simples et moins coûteuses.

Cas d'utilisation d'un LLM

Amazon Comprehend Medical peut effectuer des tâches de PNL spécifiques. Pour de plus amples informations, veuillez consulter Cas d'utilisation d'Amazon Comprehend Medical.

Les capacités d'IA logiques et génératives d'un LLM peuvent être requises pour les cas d'utilisation avancés dans le domaine des soins de santé et des sciences de la vie, tels que les suivants :

- Classification d'entités médicales personnalisées ou de catégories de texte
- Répondre aux questions cliniques
- Synthèse des rapports médicaux
- Génération et détection d'informations à partir d'informations médicales

Approches de personnalisation

Il est essentiel de comprendre comment LLMs sont mises en œuvre. LLMs sont généralement entraînés avec des milliards de paramètres, y compris des données d'entraînement provenant de nombreux domaines. Cette formation permet au LLM d'aborder les tâches les plus générales. Cependant, des défis se présentent souvent lorsque des connaissances spécifiques à un domaine sont requises. Les codes cliniques, la terminologie médicale et les informations de santé nécessaires pour générer des réponses précises sont des exemples de connaissances dans le domaine des soins de santé et des sciences de la vie. Par conséquent, l'utilisation du LLM tel quel (invite zéro sans complément de connaissance du domaine) pour ces cas d'utilisation entraîne probablement des résultats inexacts. Il existe plusieurs approches populaires que vous pouvez utiliser pour surmonter ce défi : l'ingénierie rapide, la génération augmentée par récupération (RAG) et le réglage fin.

Ingénierie rapide

L'ingénierie rapide est le processus par lequel vous orientez les solutions d'IA générative pour créer les sorties souhaitées en ajustant les entrées du LLM. En élaborant des instructions précises

avec un contexte pertinent, il est possible d'orienter le modèle vers l'exécution de tâches de santé spécialisées qui nécessitent un raisonnement. Une ingénierie rapide et efficace peut améliorer de manière significative les performances des modèles pour les cas d'utilisation dans le secteur de la santé sans nécessiter de modifications du modèle. Pour plus d'informations sur l'ingénierie rapide, consultez Implémentation de l'ingénierie rapide avancée avec Amazon Bedrock (article de AWS blog). Les instructions en quelques étapes sont des techniques que vous pouvez utiliser dans le cadre de l'ingénierie chain-of-thought rapide.

Invites avec peu d’exemples

L'invite instantanée est une technique dans laquelle vous fournissez au LLM quelques exemples des entrées-sorties souhaitées avant de lui demander d'effectuer une tâche similaire. Dans les contextes de soins de santé, cette approche est particulièrement utile pour les tâches spécialisées, telles que la reconnaissance d'entités médicales ou la synthèse de notes cliniques. En incluant 3 à 5 exemples de haute qualité dans votre message, vous pouvez améliorer de manière significative la compréhension par le modèle de la terminologie médicale et des modèles spécifiques à un domaine. Pour un exemple d'invite ponctuelle, consultez l'article de blog consacré à l'ingénierie et à la mise au point de quelques commandes dans LLMs Amazon Bedrock.AWS

Par exemple, lorsque vous extrayez des doses de médicaments à partir de notes cliniques, vous pouvez fournir des exemples de différents styles de notation qui aident le modèle à reconnaître les variations dans la façon dont les professionnels de santé documentent les prescriptions. Cette approche est particulièrement efficace lorsque vous travaillez avec des formats de documentation standardisés ou lorsque des modèles cohérents existent dans les données.

Chain-of-thought incitant

Chain-of-thought L'invite (CoT) guide le LLM dans un processus de step-by-step raisonnement. Cela le rend utile pour les tâches complexes d'aide à la décision médicale et de raisonnement diagnostique. En demandant explicitement au modèle de « réfléchir étape par étape » lors de l'analyse de scénarios cliniques, vous pouvez améliorer sa capacité à suivre les protocoles de raisonnement médical et à réduire les erreurs de diagnostic.

Cette technique excelle lorsque le raisonnement clinique nécessite plusieurs étapes logiques, telles que le diagnostic différentiel ou la planification du traitement. Cependant, cette approche présente des limites lorsqu'il s'agit de connaissances médicales hautement spécialisées autres que les données d'entraînement du modèle ou lorsqu'une précision absolue est requise pour prendre des décisions en matière de soins intensifs.

Dans ces cas, la combinaison du CoT avec une autre approche peut donner de meilleurs résultats. L'une des options consiste à combiner le CoT avec des instructions d'auto-cohérence. Pour plus d'informations, consultez Améliorer les performances des modèles de langage génératifs grâce à des instructions d'auto-cohérence sur Amazon Bedrock (AWS article de blog). Une autre option consiste à combiner des cadres de raisonnement, tels que l' ReAct invite, avec RAG. Pour plus d'informations, voir Développer des assistants avancés basés sur le chat basé sur l'IA générative à l'aide de RAG et d' ReActinstructions (directives prescriptives)AWS .

Génération à enrichissement contextuel (RAG)

La génération augmentée de récupération (RAG) est une technologie d'IA générative dans laquelle un LLM fait référence à une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation avant de générer une réponse. Un système RAG peut récupérer des informations d'ontologie médicale (telles que les classifications internationales des maladies, les fichiers nationaux sur les médicaments et les rubriques des sujets médicaux) à partir d'une source de connaissances. Cela fournit un contexte supplémentaire au LLM pour soutenir la tâche de PNL médicale.

Comme indiqué dans la Combiner Amazon Comprehend Medical avec de grands modèles linguistiques section, vous pouvez utiliser une approche RAG pour récupérer le contexte d'Amazon Comprehend Medical. Les autres sources de connaissances courantes incluent les données du domaine médical stockées dans un service de base de données, tel qu'Amazon OpenSearch Service, Amazon Kendra ou Amazon Aurora. L'extraction d'informations à partir de ces sources de connaissances peut affecter les performances de récupération, en particulier dans le cas de requêtes sémantiques utilisant une base de données vectorielle.

Une autre option pour stocker et récupérer des connaissances spécifiques à un domaine consiste à utiliser Amazon Q Business dans votre flux de travail RAG. Amazon Q Business peut indexer des référentiels de documents internes ou des sites Web destinés au public (tels que CMS.gov pour les données ICD-10). Amazon Q Business peut ensuite extraire les informations pertinentes de ces sources avant de transmettre votre requête au LLM.

Il existe plusieurs manières de créer un flux de travail RAG personnalisé. Par exemple, il existe de nombreuses manières de récupérer des données à partir d'une source de connaissances. Pour des raisons de simplicité, nous recommandons l'approche de récupération courante qui consiste à utiliser une base de données vectorielle, telle qu'Amazon OpenSearch Service, pour stocker les connaissances sous forme d'intégrations. Cela nécessite que vous utilisiez un modèle d'intégration,

tel qu'un transformateur de phrases, pour générer des intégrations pour la requête et pour les connaissances stockées dans la base de données vectorielle.

Pour plus d'informations sur les approches RAG entièrement gérées et personnalisées, consultez la section Options et architectures de génération augmentée de récupération sur. AWS

Peaufinage

Pour peaufiner un modèle existant, il faut passer un LLM, tel qu'un modèle Amazon Titan, Mistral ou Llama, puis adapter le modèle à vos données personnalisées. Il existe différentes techniques de réglage précis, dont la plupart impliquent de ne modifier que quelques paramètres au lieu de modifier tous les paramètres du modèle. C'est ce qu'on appelle le réglage fin efficace par paramètres (PEFT). Pour plus d'informations, voir Hugging Face GitHub PEFT activé.

Voici deux cas d'utilisation courants dans lesquels vous pouvez choisir d'affiner un LLM pour une tâche de PNL médicale :

- Tâche générative — Les modèles basés sur le décodeur exécutent des tâches d'IA génératives. AI/ML les praticiens utilisent des données de base pour affiner un LLM existant. Par exemple, vous pouvez former le LLM en utilisant MedQuAD, un ensemble de données public de réponses aux questions médicales. Lorsque vous appelez une requête au LLM affiné, vous n'avez pas besoin d'une approche RAG pour fournir le contexte supplémentaire au LLM.
- Embeddings — Les modèles basés sur des encodeurs génèrent des intégrations en transformant le texte en vecteurs numériques. Ces modèles basés sur des codeurs sont généralement appelés modèles d'intégration. Un modèle de transformateur de phrases est un type spécifique de modèle d'intégration optimisé pour les phrases. L'objectif est de générer des intégrations à partir du texte saisi. Les intégrations sont ensuite utilisées pour l'analyse sémantique ou pour des tâches de récupération. Pour affiner le modèle d'intégration, vous devez disposer d'un corpus de connaissances médicales, tels que des documents, que vous pouvez utiliser comme données de formation. Pour ce faire, des paires de texte basées sur la similitude ou le sentiment sont utilisées pour affiner un modèle de transformation de phrases. Pour plus d'informations, voir Entraînement et optimisation des modèles d'intégration avec Sentence Transformers v3 sur Hugging Face.

Vous pouvez utiliser Amazon SageMaker Ground Truth pour créer un ensemble de données de formation labellisé de haute qualité. Vous pouvez utiliser le jeu de données étiquetées généré par Ground Truth pour entraîner vos propres modèles. Vous pouvez également utiliser le résultat comme jeu de données d'entraînement pour un modèle Amazon SageMaker AI. Pour plus d'informations sur la reconnaissance des entités nommées, la classification du texte à étiquette unique et la

classification du texte à étiquettes multiples, consultez la section Étiquetage de texte avec Ground Truth dans la documentation Amazon SageMaker AI.

Pour plus d'informations sur le réglage précis, consultez Ajustement de grands modèles linguistiques dans le secteur de la santé ce guide.

Choisir un LLM

Amazon Bedrock est le point de départ recommandé pour évaluer les performances élevées LLMs. Pour plus d'informations, consultez la section Modèles de fondation pris en charge dans Amazon Bedrock. Vous pouvez utiliser des tâches d'évaluation de modèles dans Amazon Bedrock afin de comparer les résultats de plusieurs sorties, puis de choisir le modèle le mieux adapté à votre cas d'utilisation. Pour plus d'informations, consultez Choisir le modèle le plus performant à l'aide des évaluations Amazon Bedrock dans la documentation Amazon Bedrock.

Certains LLMs ont une formation limitée sur les données du domaine médical. Si votre cas d'utilisation nécessite de peaufiner un LLM ou un LLM qu'Amazon Bedrock ne prend pas en charge, pensez à utiliser Amazon AI. SageMaker En SageMaker IA, vous pouvez utiliser un LLM affiné ou choisir un LLM personnalisé qui a été formé sur les données du domaine médical.

Le tableau suivant répertorie les personnes les plus populaires LLMs qui ont été formées sur les données du domaine médical.

LLM	Tâches	Connaissances	Architecture
BioBert	Récupération d'informations, classification de texte et reconnaissance d'entités nommées	Résumés PubMed, articles en texte intégral et connaissances PubMedCentral générales sur le domaine	Encodeur
Clinique Albert	Récupération d'informations, classification de texte et reconnaissance d'entités nommées	Vaste ensemble de données multicentrique ainsi que plus de 3 000 000 de dossiers patients issus de systèmes de	Encodeur

		dossiers médicaux électroniques (DSE)	
GPT clinique	Récapitulatif, réponse aux questions et génération de texte	Des ensembles de données médicaux étendus et variés, y compris des dossiers médicaux, des connaissances spécifiques à un domaine et des consultations de dialogue à plusieurs niveaux	Décodeur
GatorTron-OG	Synthèse, réponse aux questions, génération de texte et recherche d'informations	Notes cliniques et littérature biomédicale	Encodeur
Med-Bert	Récupération d'informations, classification de texte et reconnaissance d'entités nommées	Vaste ensemble de données de textes médicaux, de notes cliniques, de documents de recherche et de documents liés aux soins de santé	Encodeur
Med-Palm	Réponse à des questions à des fins médicales	Ensembles de données de textes médicaux et biomédicaux	Décodeur

Alpaga médaillé	Tâches de réponse aux questions et de dialogue médical	Une variété de textes médicaux, comprenant des ressources telles que des flashcards médicaux, des wikis et des ensembles de données de dialogue	Décodeur
BioMedbert	Récupération d'informations, classification de texte et reconnaissance d'entités nommées	Exclusivement des résumés PubMed et des articles en texte intégral de PubMedCentral	Encodeur
BioMedLM	Récapitulatif, réponse aux questions et génération de texte	Littérature biomédicale issue de sources de PubMed connaissances	Décodeur

Voici les meilleures pratiques en matière d'utilisation de médecins LLMs préformés :

- Comprenez les données d'entraînement et leur pertinence pour votre tâche de PNL médicale.
- Identifiez l'architecture LLM et son objectif. Les encodeurs sont appropriés pour les intégrations et les tâches NLP. Les décodeurs sont destinés aux tâches de génération.
- Évaluez les exigences en matière d'infrastructure, de performance et de coûts pour l'hébergement du LLM médical préformé.
- Si un ajustement précis est nécessaire, assurez-vous que les données d'entraînement sont exactes sur le terrain ou que vous connaissez bien le terrain. Assurez-vous de masquer ou de supprimer les informations personnelles identifiables (PII) ou les informations de santé protégées (PHI).

Les tâches de PNL médicale dans le monde réel peuvent différer de celles d'une personne préformée LLMs en termes de connaissances ou de cas d'utilisation prévus. Si un LLM spécifique à un domaine ne répond pas à vos critères d'évaluation, vous pouvez affiner un LLM avec votre propre ensemble de données ou vous pouvez créer un nouveau modèle de base. La formation d'un nouveau modèle

de base est une entreprise ambitieuse et souvent coûteuse. Dans la plupart des cas d'utilisation, nous recommandons de peaufiner un modèle existant.

Lorsque vous utilisez ou peaufinez un LLM médical préformé, il est important de prendre en compte l'infrastructure, la sécurité et les garde-corps.

Infrastructures

Par rapport à l'utilisation d'Amazon Bedrock pour l'inférence à la demande ou par lots, l'hébergement de LLM médicaux préformés (généralement issus de Hugging Face) nécessite des ressources importantes. Pour héberger des LLM médicaux préformés, il est courant d'utiliser une image Amazon SageMaker AI qui s'exécute sur une instance Amazon Elastic Compute Cloud (Amazon EC2) avec une ou GPUs plusieurs instances, telles que des instances ml.g5 pour le calcul accéléré ou des instances ml.inf2 pour. AWS Inferentia Cela est dû au fait qu'ils LLMs consomment une grande quantité de mémoire et d'espace disque.

Sécurité et garde-corps

En fonction des exigences de conformité de votre entreprise, pensez à utiliser Amazon Comprehend et Amazon Comprehend Medical pour masquer ou supprimer les informations personnelles identifiables (PII) et les informations de santé protégées (PHI) des données de formation. Cela permet d'empêcher le LLM d'utiliser des données confidentielles lorsqu'il génère des réponses.

Nous vous recommandons de prendre en compte et d'évaluer les biais, l'équité et les hallucinations dans vos applications d'IA générative. Que vous utilisiez un LLM préexistant ou que vous le peaufiniez, mettez en place des garde-fous pour empêcher les réponses préjudiciables. Les garde-fous sont des mesures de protection que vous personnalisez en fonction des exigences de vos applications d'IA générative et de vos politiques responsables en matière d'IA. Par exemple, vous pouvez utiliser Amazon Bedrock Guardrails.

Ajustement de grands modèles linguistiques dans le secteur de la santé

L'approche d'optimisation décrite dans cette section soutient le respect des directives éthiques et réglementaires et promeut l'utilisation responsable des systèmes d'IA dans les soins de santé. Il est conçu pour générer des informations précises et confidentielles. L'IA générative révolutionne la prestation des soins de santé, mais les off-the-shelf modèles sont souvent insuffisants dans les environnements cliniques où la précision est essentielle et où la conformité n'est pas négociable.

L'affinement des modèles de base à l'aide de données spécifiques au domaine comble cette lacune. Il vous aide à créer des systèmes d'IA qui parlent le langage de la médecine tout en respectant des normes réglementaires strictes. Cependant, pour réussir un ajustement précis, il faut relever avec soin les défis uniques des soins de santé : protéger les données sensibles, justifier les investissements dans l'IA par des résultats mesurables et maintenir la pertinence clinique dans des environnements médicaux en évolution rapide.

Lorsque les approches plus légères atteignent leurs limites, le peaufinage devient un investissement stratégique. On s'attend à ce que les gains de précision, de latence ou d'efficacité opérationnelle compensent les coûts de calcul et d'ingénierie importants nécessaires. Il est important de se rappeler que le rythme de progression des modèles de base est rapide, de sorte que l'avantage d'un modèle affiné peut ne durer que jusqu'à la prochaine sortie majeure du modèle.

Cette section ancre la discussion dans les deux cas d'utilisation à fort impact suivants, réalisés par des clients du AWS secteur de la santé :

- Systèmes d'aide à la décision clinique — Améliorez la précision des diagnostics grâce à des modèles qui comprennent les antécédents complexes des patients et l'évolution des directives. Un ajustement précis peut aider les modèles à comprendre en profondeur les antécédents complexes des patients et à intégrer des directives spécialisées, ce qui peut potentiellement réduire les erreurs de prédiction des modèles. Cependant, vous devez évaluer ces gains par rapport au coût de la formation sur de grands ensembles de données sensibles et à l'infrastructure requise pour les applications cliniques à enjeux élevés. L'amélioration de la précision et de la connaissance du contexte justifiera-t-elle l'investissement, en particulier lorsque de nouveaux modèles sont publiés fréquemment ?
- Analyse des documents médicaux — Automatisez le traitement des notes cliniques, des rapports d'imagerie et des documents d'assurance tout en respectant la loi HIPAA (Health Insurance Portability and Accountability Act). Dans ce cas, un ajustement précis peut permettre au modèle de gérer plus efficacement les formats uniques, les abréviations spécialisées et les exigences réglementaires. Les avantages se traduisent souvent par une réduction du temps de révision manuelle et une meilleure conformité. Néanmoins, il est essentiel d'évaluer si ces améliorations sont suffisamment importantes pour justifier le réglage précis des ressources. Déterminez si une ingénierie et une orchestration rapides des flux de travail peuvent répondre à vos besoins.

Ces scénarios concrets illustrent le processus de mise au point, de l'expérimentation initiale au déploiement du modèle, tout en répondant aux exigences uniques des soins de santé à chaque étape.

Estimation des coûts et du retour sur investissement

Les facteurs de coût suivants doivent être pris en compte lors de la mise au point d'un LLM :

- Taille du modèle — Les modèles plus grands coûtent plus cher à peaufiner
- Taille du jeu de données — Les coûts et le temps de calcul augmentent avec la taille du jeu de données pour un ajustement précis
- Stratégie de réglage précis — Les méthodes efficaces en termes de paramètres peuvent réduire les coûts par rapport aux mises à jour complètes des paramètres

Lorsque vous calculez le retour sur investissement (ROI), considérez l'amélioration des indicateurs que vous avez choisis (tels que la précision) multipliée par le volume de demandes (fréquence d'utilisation du modèle) et la durée attendue avant que le modèle ne soit dépassé par les nouvelles versions.

Tenez également compte de la durée de vie de votre LLM de base. De nouveaux modèles de base apparaissent tous les 6 à 12 mois. S'il faut 8 mois pour peaufiner et valider votre détecteur de maladies rares, il se peut que vous n'obteniez que 4 mois de performances supérieures avant que les nouveaux modèles ne comblent l'écart.

En calculant les coûts, le retour sur investissement et la durée de vie potentielle de votre cas d'utilisation, vous pouvez prendre une décision basée sur les données. Par exemple, si le peaufinage de votre modèle d'aide à la décision clinique entraîne une réduction mesurable des erreurs de diagnostic dans des milliers de cas par an, l'investissement pourrait rapidement porter ses fruits. À l'inverse, si une ingénierie rapide permet à elle seule de rapprocher votre flux de travail d'analyse de documents de la précision cible, il peut être judicieux de ne pas peaufiner les réglages jusqu'à l'arrivée de la prochaine génération de modèles.

Le réglage fin ne l'est pas one-size-fits-all. Si vous décidez de peaufiner, la bonne approche dépend de votre cas d'utilisation, de vos données et de vos ressources.

Choisir une stratégie de réglage précis

Une fois que vous avez déterminé que le réglage précis est la bonne approche pour votre cas d'utilisation dans le secteur de la santé, l'étape suivante consiste à sélectionner la stratégie d'ajustement la plus appropriée. Plusieurs approches sont disponibles. Chacune présente des avantages et des inconvénients distincts pour les applications de santé. Le choix entre ces méthodes

dépend de vos objectifs spécifiques, des données disponibles et des contraintes en matière de ressources.

Objectifs de formation

Le pré-entraînement adaptatif au domaine (DAPT) est une méthode non supervisée qui consiste à pré-entraîner le modèle sur un grand nombre de textes non étiquetés spécifiques au domaine (tels que des millions de documents médicaux). Cette approche convient parfaitement pour améliorer la capacité des modèles à comprendre les abréviations des spécialités médicales et la terminologie utilisée par les radiologues, les neurologues et les autres prestataires spécialisés. Cependant, DAPT nécessite de grandes quantités de données et ne traite pas de résultats de tâches spécifiques.

Le réglage fin supervisé (SFT) apprend au modèle à suivre des instructions explicites en utilisant des exemples d'entrées-sorties structurés. Cette approche excelle pour les flux de travail d'analyse de documents médicaux, tels que le résumé de documents ou le codage clinique. Le réglage des instructions est une forme courante de SFT dans laquelle le modèle est entraîné sur des exemples qui incluent des instructions explicites associées aux sorties souhaitées. Cela améliore la capacité du modèle à comprendre et à suivre les diverses instructions des utilisateurs. Cette technique est particulièrement utile dans les établissements de santé car elle entraîne le modèle à l'aide d'exemples cliniques spécifiques. Le principal inconvénient est qu'il nécessite des exemples soigneusement étiquetés. En outre, le modèle affiné peut avoir du mal à traiter les cas extrêmes où il n'y a pas d'exemples. Pour obtenir des instructions sur le réglage précis avec Amazon SageMaker Jumpstart, consultez la section Instructions de réglage du FLAN T5 XL avec Amazon SageMaker Jumpstart (article de blog).AWS

L'apprentissage par renforcement basé sur le feedback humain (RLHF) optimise le comportement du modèle en fonction des commentaires et des préférences des experts. Utilisez un modèle de récompense basé sur les préférences et méthodes humaines, telles que l'optimisation des politiques proximales (PPO) ou l'optimisation des préférences directes (DPO), pour optimiser le modèle tout en empêchant les mises à jour destructives. Le RLHF est idéal pour aligner les résultats sur les directives cliniques et s'assurer que les recommandations respectent les protocoles approuvés. Cette approche demande beaucoup de temps aux cliniciens pour obtenir des commentaires et implique un pipeline de formation complexe. Cependant, le RLHF est particulièrement utile dans le domaine de la santé car il aide les experts médicaux à façonner la manière dont les systèmes d'IA communiquent et font des recommandations. Par exemple, les cliniciens peuvent fournir des commentaires pour s'assurer que le modèle fonctionne correctement au chevet du patient, qu'il sait quand exprimer son incertitude et qu'il respecte les directives cliniques. Des techniques telles que le PPO optimisent de manière itérative le comportement du modèle en fonction des commentaires des experts tout en

limitant les mises à jour des paramètres afin de préserver les connaissances médicales de base. Cela permet aux modèles de transmettre des diagnostics complexes dans un langage convivial pour le patient tout en signalant les affections graves nécessitant une prise en charge médicale immédiate. Cela est crucial pour les soins de santé où la précision et le style de communication sont importants. Pour plus d'informations sur le RLHF, voir Affiner les grands modèles linguistiques grâce à l'apprentissage par renforcement à partir de commentaires humains ou basés sur l'IA (article de AWS blog).

Méthodes de mise en œuvre

Une mise à jour complète des paramètres implique la mise à jour de tous les paramètres du modèle pendant l'entraînement. Cette approche fonctionne mieux pour les systèmes d'aide à la décision clinique qui nécessitent une intégration approfondie des antécédents des patients, des résultats de laboratoire et des directives évolutives. Les inconvénients incluent le coût de calcul élevé et le risque de surajustement si votre ensemble de données n'est pas volumineux et diversifié.

Les méthodes de réglage fin efficaces (PEFT) mettent à jour uniquement un sous-ensemble de paramètres afin d'éviter un surajustement ou une perte catastrophique des capacités linguistiques. Les types incluent l'adaptation de bas rang (LoRa), les adaptateurs et le réglage des préfixes. Les méthodes PEFT permettent de réduire les coûts de calcul, d'accélérer la formation et sont idéales pour les expériences telles que l'adaptation d'un modèle d'aide à la décision clinique aux protocoles ou à la terminologie d'un nouvel hôpital. La principale limite est la réduction potentielle des performances par rapport aux mises à jour complètes des paramètres.

Pour plus d'informations sur les méthodes de réglage précis, consultez Méthodes de réglage avancées sur Amazon SageMaker AI (article de AWS blog).

Création d'un ensemble de données affiné

La qualité et la diversité de l'ensemble de données de réglage fin sont essentielles pour les performances, la sécurité et la prévention des biais du modèle. Les trois domaines critiques suivants doivent être pris en compte lors de la création de cet ensemble de données :

- Volume basé sur une approche de réglage précis
- Annotation des données par un expert du domaine
- Diversité de l'ensemble de données

Comme le montre le tableau suivant, les exigences relatives à la taille du jeu de données pour le réglage précis varient en fonction du type de réglage précis effectué.

Stratégie de réglage précis	Taille du jeu de données
Pré-formation adaptée au domaine	Plus de 100 000 textes de domaine
Réglage précis supervisé	Plus de 10 000 paires étiquetées
Apprentissage par renforcement à partir du feedback humain	Plus de 1 000 paires de préférences d'experts

Vous pouvez utiliser AWS GlueAmazon EMR et Amazon SageMaker Data Wrangler pour automatiser le processus d'extraction et de transformation des données afin de créer un ensemble de données dont vous êtes le propriétaire. Si vous ne parvenez pas à créer un ensemble de données suffisamment volumineux, vous pouvez découvrir et télécharger des ensembles de données directement dans votre Compte AWS canal. AWS Data Exchange Consultez votre conseiller juridique avant d'utiliser des ensembles de données tiers.

Des annotateurs experts ayant une connaissance du domaine, tels que les médecins, les biologistes et les chimistes, devraient participer au processus de curation des données afin d'intégrer les nuances des données médicales et biologiques dans les résultats du modèle. Amazon SageMaker Ground Truth fournit une interface utilisateur low-code permettant aux experts d'annoter l'ensemble de données.

Un ensemble de données représentant la population humaine est essentiel pour que les soins de santé et les sciences de la vie puissent affiner les cas d'utilisation afin d'éviter les biais et de refléter les résultats du monde réel.AWS Glue les sessions interactives ou les instances de SageMaker blocs-notes Amazon constituent un moyen puissant d'explorer de manière itérative des ensembles de données et d'affiner les transformations à l'aide de blocs-notes compatibles avec Jupyter. Les sessions interactives vous permettent de travailler avec un choix d'environnements de développement intégrés populaires (IDEs) dans votre environnement local. Vous pouvez également travailler avec AWS Glue des blocs-notes Amazon SageMaker Studio via le AWS Management Console.

Affiner le modèle

AWS fournit des services tels qu'Amazon SageMaker AI et Amazon Bedrock qui sont essentiels pour un réglage précis réussi.

SageMaker L'IA est un service d'apprentissage automatique entièrement géré qui aide les développeurs et les data scientists à créer, former et déployer rapidement des modèles de machine learning. Les trois fonctionnalités utiles de l' SageMaker IA pour le peaufinage sont les suivantes :

- SageMakerFormation — Une fonctionnalité de machine learning entièrement gérée qui vous aide à entraîner efficacement un large éventail de modèles à grande échelle
- SageMaker JumpStart— Une fonctionnalité qui s'appuie sur les tâches de SageMaker formation pour fournir des modèles préentraînés, des algorithmes intégrés et des modèles de solutions pour les tâches de machine learning
- SageMaker HyperPod— Une solution d'infrastructure spécialement conçue pour la formation distribuée des modèles de base et LLMs

Amazon Bedrock est un service entièrement géré qui donne accès à des modèles de base très performants via une API, avec des fonctionnalités intégrées de sécurité, de confidentialité et d'évolutivité. Le service permet de peaufiner plusieurs modèles de base disponibles. Pour plus d'informations, consultez la section Modèles et régions pris en charge pour un réglage précis et une formation préalable continue dans la documentation Amazon Bedrock.

Lorsque vous abordez le processus de mise au point avec l'un ou l'autre service, tenez compte du modèle de base, de la stratégie de réglage et de l'infrastructure.

Choix du modèle de base

Les modèles à code source fermé, tels qu'Anthropic Claude, Meta Llama et Amazon Nova, offrent de solides out-of-the-box performances grâce à la gestion de la conformité, mais limitent la flexibilité de réglage aux options prises en charge par les fournisseurs, telles que celles gérées par Amazon Bedrock. APIs Cela limite la personnalisation, en particulier pour les cas d'utilisation des soins de santé réglementés. En revanche, les modèles open source, tels que Meta Llama, offrent un contrôle et une flexibilité complets sur l'ensemble des services Amazon SageMaker AI, ce qui les rend idéaux lorsque vous devez personnaliser, auditer ou adapter en profondeur un modèle à vos exigences spécifiques en matière de données ou de flux de travail.

Stratégie de peaufinage

Le réglage simple des instructions peut être géré par Amazon Bedrock Model Customization ou Amazon SageMaker JumpStart. Les approches PEFT complexes, telles que LoRa ou les adaptateurs, nécessitent des tâches de SageMaker formation ou une fonctionnalité de réglage personnalisé dans Amazon Bedrock. La formation distribuée pour les très grands modèles est prise en charge par SageMaker HyperPod.

Échelle et contrôle de l'infrastructure

Les services entièrement gérés, tels qu'Amazon Bedrock, minimisent la gestion de l'infrastructure et sont idéaux pour les entreprises qui privilégient la facilité d'utilisation et la conformité. Les options semi-gérées, par exemple SageMaker JumpStart, offrent une certaine flexibilité avec moins de complexité. Ces options conviennent au prototypage rapide ou à l'utilisation de flux de travail prédéfinis. Les tâches de SageMaker formation offrent un contrôle et une personnalisation complets HyperPod, bien que celles-ci nécessitent une plus grande expertise et soient idéales lorsque vous devez effectuer une mise à l'échelle pour des ensembles de données volumineux ou que vous avez besoin de pipelines personnalisés.

Surveillance de modèles affinés

Dans les domaines de la santé et des sciences de la vie, le suivi de l'ajustement précis du LLM nécessite le suivi de plusieurs indicateurs de performance clés. La précision fournit une mesure de référence, mais cela doit être mis en balance avec la précision et le rappel, en particulier dans les applications où les erreurs de classification ont des conséquences importantes. Le score F1 aide à résoudre les problèmes de déséquilibre des classes qui peuvent être courants dans les ensembles de données médicales. Pour plus d’informations, consultez Évaluation LLMs pour les applications des soins de santé et des sciences de la vie dans ce guide.

Les mesures d'étalonnage vous aident à vous assurer que les niveaux de confiance du modèle correspondent aux probabilités réelles. Les indicateurs d'équité peuvent vous aider à détecter les biais potentiels selon les différents groupes démographiques de patients.

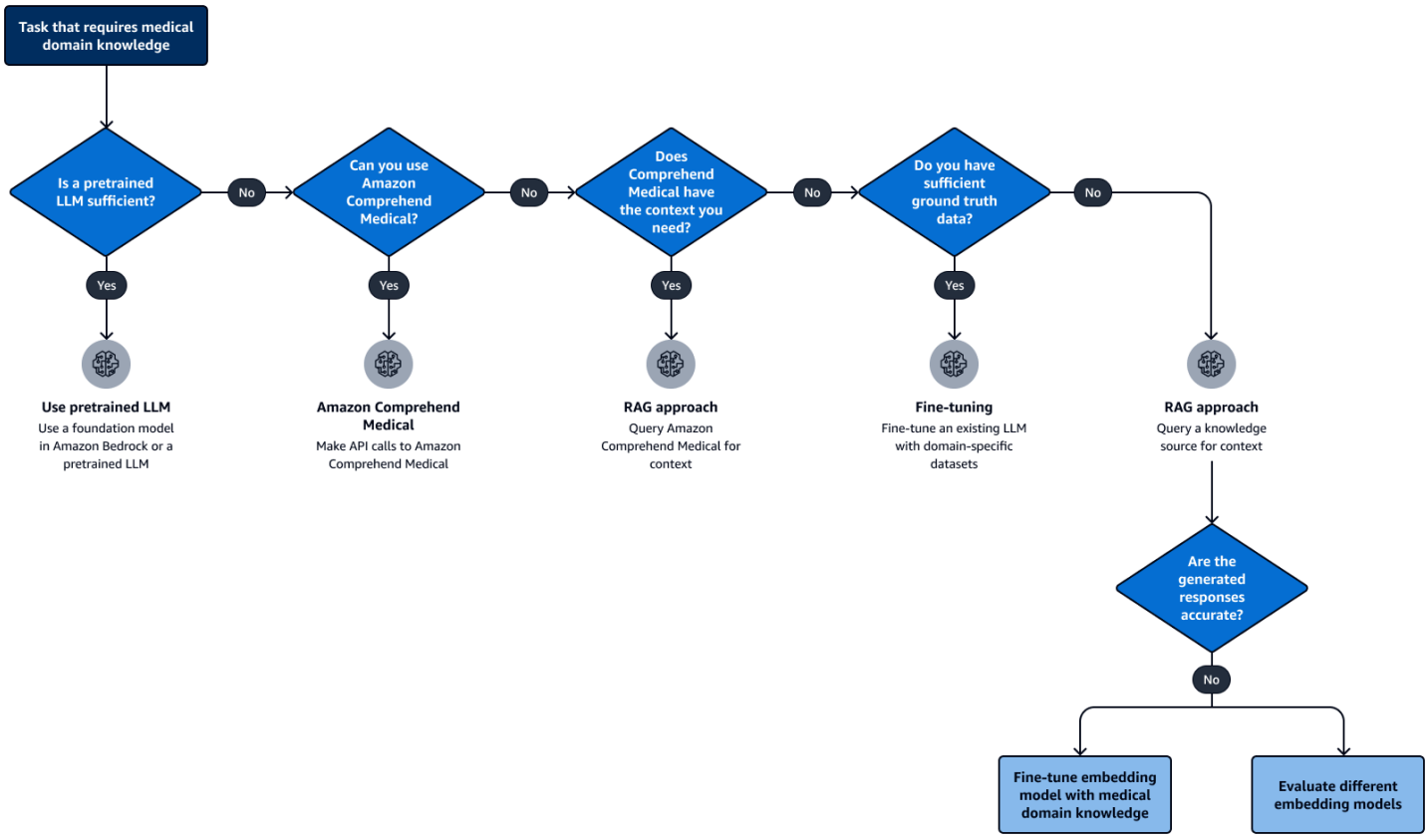
MLflowest une solution open source qui peut vous aider à suivre des expériences de réglage précis. MLflow est pris en charge de manière native dans Amazon SageMaker AI, ce qui vous permet de comparer visuellement les indicateurs issus des sessions d'entraînement. Pour affiner les tâches sur Amazon Bedrock, les statistiques sont transmises à Amazon CloudWatch afin que vous puissiez les visualiser dans la console. CloudWatch

Choisir une approche de PNL pour les soins de santé et les sciences de la vie

La Approches génératives de l'IA et de la PNL pour les soins de santé et les sciences de la vie section décrit les approches suivantes pour traiter les tâches de traitement du langage naturel (NLP) pour les applications des soins de santé et des sciences de la vie :

- Utilisation d'Amazon Comprehend Medical
- Combiner Amazon Comprehend Medical avec un LLM dans un flux de travail RAG (Retrieval Augment Generation)
- Utilisation d'un LLM affiné
- Utilisation d'un flux de travail RAG

En évaluant les limites connues LLMs des tâches du domaine médical et votre cas d'utilisation, vous pouvez choisir l'approche la mieux adaptée à votre tâche. L'arbre de décision suivant peut vous aider à choisir une approche LLM pour votre tâche de PNL médicale :



Le schéma suivant illustre le flux de travail suivant :

1. Pour les cas d'utilisation dans le domaine des soins de santé et des sciences de la vie, déterminez si la tâche de PNL nécessite des connaissances spécifiques dans le domaine. Au besoin, coordonnez-vous avec les experts en la matière (SMEs).
2. Si vous pouvez utiliser un LLM général ou un modèle formé sur des ensembles de données médicaux, utilisez un modèle de base disponible dans Amazon Bedrock ou le LLM préentraîné. Pour plus d'informations, consultez Choisir un LLM dans ce guide.
3. Si les fonctionnalités de détection d'entités et de liaison d'ontologies d'Amazon Comprehend Medical répondent à votre cas d'utilisation, utilisez Amazon Comprehend Medical. APIs Pour plus d'informations, consultez Utilisation d'Amazon Comprehend Medical dans ce guide.
4. Parfois, Amazon Comprehend Medical dispose du contexte requis mais ne prend pas en charge votre cas d'utilisation. Par exemple, vous pourriez avoir besoin de définitions d'entités différentes, recevoir un très grand nombre de résultats, avoir besoin d'entités personnalisées ou avoir besoin d'une tâche NLP personnalisée. Si tel est le cas, utilisez une approche RAG pour interroger Amazon Comprehend Medical pour obtenir le contexte. Pour plus d'informations, consultez Combiner Amazon Comprehend Medical avec de grands modèles linguistiques dans ce guide.
5. Si vous disposez d'une quantité suffisante de données fiables, affinez un LLM existant. Pour plus d'informations, consultez Approches de personnalisation dans ce guide.
6. Si les autres approches ne répondent pas aux objectifs médicaux de vos tâches de PNL, implémentez une solution RAG. Pour plus d'informations, consultez Approches de personnalisation dans ce guide.
7. Après avoir implémenté la solution RAG, évaluez si les réponses générées sont exactes. Pour plus d'informations, consultez Évaluation LLMs pour les applications des soins de santé et des sciences de la vie dans ce guide. Il est courant de commencer par un modèle Amazon Titan Text Embeddings ou un modèle de transformateur de phrases générales, tel que All-MiniLM-L6-V2. Cependant, en raison de l'absence de contexte du domaine, ces modèles peuvent ne pas saisir la terminologie médicale du texte. Si nécessaire, pensez aux ajustements suivants :
   a. Évaluer d'autres modèles d'intégration
   b. Affinez le modèle d'intégration avec des ensembles de données spécifiques au domaine

Considérations relatives à la maturité commerciale

La maturité commerciale est essentielle lors de l'adaptation des solutions LLM pour les applications de santé et des sciences de la vie. Ces organisations sont confrontées à différents niveaux de complexité lors de la mise en œuvre LLMs, en fonction de leurs critères d'acceptation. Souvent, les organisations qui manquent de AI/ML ressources investissent dans le soutien de sous-traitants pour créer des solutions LLM. Dans ces situations, il est important de comprendre les compromis suivants :

- Des performances élevées pour des coûts et une maintenance élevés — Vous pourriez avoir besoin d'une solution complexe nécessitant des ajustements ou une personnalisation LLMs pour répondre à des normes de performance strictes. Cependant, cela entraîne des coûts et des exigences de maintenance plus élevés. Vous devrez peut-être engager des ressources spécialisées ou vous associer à des sous-traitants pour maintenir ces solutions sophistiquées. Cela peut potentiellement ralentir le développement.
- De bonnes performances pour un faible coût et une maintenance réduits. Vous pourriez également constater que des services tels qu'Amazon Bedrock ou Amazon Comprehend Medical offrent des performances acceptables. Bien que ces approches LLMs ou ces approches puissent fournir des résultats parfaits, ces solutions peuvent souvent fournir des résultats cohérents et de haute qualité. Ces solutions sont moins coûteuses et réduisent la charge de maintenance. Cela peut accélérer le développement.

Si une approche plus simple et moins coûteuse produit systématiquement des résultats de haute qualité qui répondent à vos critères d'acceptation, demandez-vous si l'augmentation des performances vaut les compromis en termes de coûts, de maintenance et de temps. Toutefois, si la solution la plus simple est nettement inférieure aux performances cibles, et si votre organisation ne dispose pas de la capacité d'investissement nécessaire pour les solutions complexes et leurs exigences de maintenance, envisagez de reporter le AI/ML développement jusqu'à ce que davantage de ressources ou des solutions alternatives soient disponibles.

En outre, pour toute solution de PNL médicale qui repose sur un LLM, nous vous recommandons d'effectuer un suivi et une évaluation continus. Évaluez les commentaires des utilisateurs au fil du temps et mettez en œuvre des évaluations périodiques pour vous assurer que la solution continue de répondre à vos objectifs commerciaux.

Évaluation LLMs pour les applications des soins de santé et des sciences de la vie

Cette section fournit un aperçu complet des exigences et des considérations relatives à l'évaluation de grands modèles linguistiques (LLMs) dans les cas d'utilisation des soins de santé et des sciences de la vie.

Il est important d'utiliser des données fiables sur le terrain et les commentaires des PME pour atténuer les biais et valider l'exactitude de la réponse générée par le LLM. Cette section décrit les meilleures pratiques en matière de collecte et de conservation des données de formation et de test. Il vous aide également à mettre en place des garde-fous et à mesurer le biais et l'équité des données. Il aborde également les tâches médicales courantes de traitement du langage naturel (NLP), telles que la classification de texte, la reconnaissance d'entités nommées et la génération de texte, ainsi que les mesures d'évaluation associées.

Il présente également des flux de travail pour effectuer l'évaluation du LLM pendant la phase d'expérimentation de la formation et la phase de post-production. Le suivi des modèles et les opérations de LLM sont des éléments importants de ce processus d'évaluation.

Données de formation et de test pour les tâches de PNL médicale

Les tâches de PNL médicale utilisent généralement des corpus médicaux (tels que PubMed) ou des informations sur les patients (telles que les notes de visite des patients en clinique) pour classer, résumer et générer des informations. Le personnel médical, tel que les médecins, les administrateurs de soins de santé ou les techniciens, varie en termes d'expertise et de points de vue. En raison de la subjectivité entre ces personnels médicaux, des ensembles de données de formation et de tests plus restreints présentent un risque de biais. Pour atténuer ce risque, nous recommandons les meilleures pratiques suivantes :

- Lorsque vous utilisez une solution LLM préentraînée, assurez-vous de disposer d'une quantité suffisante de données de test. Les données du test doivent ressembler étroitement aux données médicales réelles. Selon la tâche, cela peut aller de 20 à plus de 100 enregistrements.
- Lorsque vous peaufinez un LLM, collectez un nombre suffisant de dossiers étiquetés (Ground Truth) provenant SMEs de divers domaines médicaux ciblés. Le point de départ général est d'avoir au moins 100 enregistrements de haute qualité. Toutefois, compte tenu de la complexité

de la tâche et de vos critères d'acceptation de la précision, d'autres enregistrements peuvent être nécessaires.

- Si cela est nécessaire pour votre cas d'utilisation médicale, mettez en place des garde-fous et mesurez le biais et l'équité des données. Par exemple, assurez-vous que le LLM prévient les erreurs de diagnostic dues au profil racial des patients. Pour plus d'informations, consultez la Sécurité et garde-corps section de ce guide.

De nombreuses sociétés de recherche et développement dans le domaine de l'IA, comme Anthropic, ont déjà intégré des garde-fous dans leurs modèles de base afin d'éviter toute toxicité. Vous pouvez utiliser la détection de toxicité pour vérifier les instructions d'entrée et les réponses de sortie de LLMs. Pour plus d'informations, consultez la section Détection de toxicité dans la documentation Amazon Comprehend et consultez Guardrails dans la documentation Amazon Bedrock.

Toute tâche d'IA générative comporte un risque d'hallucination. Vous pouvez atténuer ce risque en effectuant des tâches de PNL, telles que la classification. Vous pouvez également utiliser des techniques plus avancées, telles que les mesures de similarité de texte. BertScoreest une métrique de similarité de texte couramment adoptée. Pour plus d'informations sur les techniques que vous pouvez utiliser pour atténuer les hallucinations, voir Une enquête complète sur les techniques d'atténuation des hallucinations dans les grands modèles linguistiques.

Indicateurs pour les tâches de PNL médicale

Vous pouvez créer des mesures quantifiables après avoir établi des données fiables sur le terrain et des étiquettes fournies par les PME pour la formation et les tests. Le contrôle de la qualité par le biais de processus qualitatifs, tels que les tests de stress et l'examen des résultats du LLM, est utile pour un développement rapide. Cependant, les métriques agissent comme des repères quantitatifs qui soutiennent les futures opérations de LLM et servent de repères de performance pour chaque version de production.

Il est essentiel de comprendre la tâche médicale. Les métriques correspondent généralement à l'une des tâches générales de PNL suivantes :

- Classification du texte — Le LLM classe le texte dans une ou plusieurs catégories prédéfinies, en fonction de l'invite de saisie et du contexte fourni. Par exemple, on peut classer une catégorie de douleur à l'aide d'une échelle de douleur. Voici des exemples de mesures de classification de texte :
  - Précision

  - Précision, également connue sous le nom de macroprécision
  - Rappel, également connu sous le nom de rappel de macros
  - Score F1, également connu sous le nom de score F1 macro
  - Défaite de Hamming
- Reconnaissance d'entités nommées (NER) — Également connue sous le nom d'extraction de texte, la reconnaissance d'entités nommées est le processus de localisation et de classification des entités nommées mentionnées dans un texte non structuré dans des catégories prédéfinies. L'extraction des noms des médicaments des dossiers des patients en est un exemple. Voici des exemples de métriques NER :
  - Précision
  - Précision
  - Rappel
  - Score de F1
  - Défaite de Hamming
- Génération — Le LLM génère un nouveau texte en traitant l'invite et le contexte fourni. La génération inclut des tâches de synthèse ou des tâches de réponse à des questions. Voici des exemples de mesures de génération :
  - Doublure axée sur les rappels pour l'évaluation du gisting (ROUGE)
  - Métrique pour l'évaluation de la traduction avec Explicit ORdering (METEOR)
  - Doublure d'évaluation bilingue (BLEU) (pour les traductions)
  - Distance entre chaînes, également connue sous le nom de similitude en cosinus

FAQ sur les cas d'utilisation des soins de santé et des sciences de la vie

Les questions suivantes sont fréquemment posées concernant l'utilisation d'Amazon Comprehend Medical LLMs ou les tâches de PNL médicale.

Comment choisir entre Amazon Comprehend Medical et un LLM ?

Si votre tâche consiste à détecter des entités médicales dans votre texte médical, consultez la documentation Amazon Comprehend Medical pour savoir quelles entités médicales peuvent être extraites et si l'une des ontologies répond à votre cas d'utilisation. Si ce n'est pas le cas, envisagez d'utiliser un LLM. Pour plus d'informations, consultez Cas d'utilisation d'Amazon Comprehend Medical et Cas d'utilisation d'un LLM dans ce guide.

Comment puis-je fournir les résultats d'Amazon Comprehend Medical à un LLM ?

Vous pouvez intégrer les résultats d'Amazon Comprehend Medical en tant que contexte dans vos instructions de maîtrise en droit. Cela fournit des connaissances et une terminologie médicales supplémentaires au LLM. Le contexte fourni peut améliorer les performances du LLM sur des tâches telles que la reconnaissance d'entités, la synthèse ou la réponse aux questions. Le guide fournit plusieurs exemples de la manière de structurer les invites à l'aide des résultats d'Amazon Comprehend Medical. Pour plus d’informations, consultez Combiner Amazon Comprehend Medical avec de grands modèles linguistiques dans ce guide.

Quelles sont les meilleures pratiques à suivre lors de l'utilisation d'Amazon Comprehend Medical LLMs avec ?

Nous vous recommandons d'utiliser les scores de confiance d'Amazon Comprehend Medical pour filtrer ou hiérarchiser les entités figurant dans vos instructions. Il est également important d'évaluer ses performances sur vos données spécifiques et de vérifier que les définitions des entités correspondent à vos exigences. La combinaison d'Amazon Comprehend Medical avec des sources de connaissances spécifiques à un domaine peut encore améliorer les performances du LLM. Pour

plus d'informations, consultez Bonnes pratiques pour utiliser Amazon Comprehend Medical dans un flux de travail RAG dans ce guide.

Dois-je utiliser un LLM médical préformé ou peaufiner un LLM général pour mon cas d'utilisation dans le secteur de la santé ?

La décision dépend de vos besoins spécifiques et de la disponibilité de données d'entraînement de haute qualité. Un médecin préformé LLMs peut constituer un bon point de départ. Toutefois, il se peut que vous deviez les ajuster avec les données spécifiques à votre domaine. Si vous disposez de suffisamment de données étiquetées, le réglage précis d'un LLM général peut être une option viable. Pour plus d'informations, consultez Choisir un LLM et Choisir une approche de PNL pour les soins de santé et les sciences de la vie dans ce guide.

Comment puis-je évaluer les performances des tâches LLMs de PNL médicale ?

Nous recommandons d'utiliser des mesures quantitatives, telles que l'exactitude, la précision, le rappel et le score F1 pour la classification du texte et les tâches de reconnaissance d'entités nommées. Vous pouvez utiliser ROUGE et METEOR pour les tâches de génération de texte. Il est important de disposer de données fiables certifiées par des experts en la matière et de mettre en œuvre des processus de surveillance des performances des modèles au fil du temps. Pour plus d'informations, consultez Évaluation LLMs pour les applications des soins de santé et des sciences de la vie dans ce guide.

Quels sont les compromis entre les solutions LLM très complexes et les solutions LLM peu complexes ?

La mise au point d'un LLM ou la création d'un LLM personnalisé sont des solutions très complexes. Ces approches peuvent améliorer les performances, mais elles entraînent des coûts et des exigences de maintenance plus élevés. Des solutions plus simples, telles que l'utilisation d'Amazon préformé LLMs ou d'Amazon Comprehend Medical, peuvent fournir des performances acceptables avec des coûts réduits et des cycles de développement plus rapides. Cependant, ces approches peuvent ne pas répondre à des exigences de précision strictes dans certains cas d'utilisation. Pour plus d'informations, consultez Considérations relatives à la maturité commerciale dans ce guide.

Prochaines étapes et ressources

Ce guide vous aide Services AWS à automatiser les tâches de PNL médicale et d'IA générative pour des applications réelles dans des environnements de production. Il décrit comment vous pouvez utiliser Amazon Comprehend Medical, LLMs pris en charge par Amazon Bedrock, préformé dans le domaine LLMs médical ou LLMs peaufiné pour atteindre vos objectifs commerciaux dans le domaine de la santé et des sciences de la vie. Ce guide décrit les avantages et les limites des approches suivantes :

- Utiliser Amazon Comprehend Medical de manière indépendante
- Fournir les résultats d'Amazon Comprehend Medical à un LLM
- Utilisation d'un LLM général préformé ou d'un LLM médical dans une approche de génération augmentée par récupération (RAG)
- Perfectionnement d'un LLM général ou d'un LLM médical

Utilisez l'arbre décisionnel et les considérations relatives à la maturité commerciale de ce guide pour choisir entre ces approches en fonction du niveau de AI/ML maturité de votre organisation. Bien qu'Amazon Comprehend Medical et Amazon LLMs Bedrock proposent de puissantes fonctionnalités, elles ne sont efficaces que si vous les implémentez et les évaluez correctement. Utilisez les informations d'évaluation et les mesures décrites dans ce guide pour valider les performances de votre solution.

Pour les prochaines étapes, nous recommandons aux responsables informatiques, aux architectes et aux responsables techniques du secteur de la santé de travailler avec les AI/ML praticiens pour identifier leur tâche médicale en matière de PNL. Utilisez ce guide pour choisir un chemin de développement, puis utilisez les fonctionnalités appropriées Services AWS pour implémenter avec succès une solution automatisée sur AWS.

AWS ressources

- Documentation Amazon Comprehend Medical :
  - Manuel du développeur
  - API Reference
- Documentation Amazon Bedrock

  - Évaluation du modèle Amazon Bedrock
  - Réglage précis dans Amazon Bedrock
- Affiner un modèle dans Amazon AI SageMaker
- Amazon SageMaker Ground Truth
- Détection de toxicité par Amazon Comprehend
- AWS Partenaires spécialisés dans le domaine de la santé

Autres ressources

- Classement Open Medical-LLM
- Une enquête sur les grands modèles linguistiques pour les soins de santé : des données, de la technologie et des applications à la responsabilité et à l'éthique
- Les grands modèles linguistiques sont de mauvais codeurs médicaux — Analyse comparative des requêtes de code médical
- Du débutant à l'expert : transformer les connaissances médicales en connaissances générales LLMs

Collaborateurs

Conception

- Joe King, scientifique AWS principal des données
- Ankith Ede, architecte de AWS solutions
- Clément Perrot, stratège AWS principal en IA générative
- Jillian Forde, architecte de solutions AWS senior
- Rajesh Sitaraman, consultant principal en livraison AWS
- Ross Claytor, chercheur appliqué AWS principal
- Shivesh Ummat, architecte de solutions AWS

Révision

- Dilshad Raihan Akkam Veettil, data scientist senior AWS
- Joseph Cottingham, architecte du AWS Deep Learning

Rédaction technique

- Lilly AbouHarb, AWS rédactrice technique principale

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
Nouvelles sections	Nous avons ajouté la section Affiner les grands modèles linguistiques dans les soins de santé et la section Prompt engineering.	5 décembre 2025
Publication initiale	—	16 décembre 2024

AWS Glossaire des directives prescriptives

Les termes suivants sont couramment utilisés dans les stratégies, les guides et les modèles fournis par les directives AWS prescriptives. Pour suggérer des entrées, veuillez utiliser le lien Faire un commentaire à la fin du glossaire.

Nombres

7 R

Sept politiques de migration courantes pour transférer des applications vers le cloud. Ces politiques s'appuient sur les 5 R identifiés par Gartner en 2011 et sont les suivantes :

- Refactorisation/réarchitecture : transférez une application et modifiez son architecture en tirant pleinement parti des fonctionnalités natives cloud pour améliorer l'agilité, les performances et la capacité de mise à l'échelle. Cela implique généralement le transfert du système d'exploitation et de la base de données. Exemple : migrez votre base de données Oracle sur site vers l'édition compatible avec Amazon Aurora PostgreSQL.
- Replateformer (déplacer et remodeler) : transférez une application vers le cloud et introduisez un certain niveau d'optimisation pour tirer parti des fonctionnalités du cloud. Exemple : migrez votre base de données Oracle sur site vers Amazon Relational Database Service (Amazon RDS) pour Oracle dans le. AWS Cloud
- Racheter (rachat) : optez pour un autre produit, généralement en passant d'une licence traditionnelle à un modèle SaaS. Exemple : migrez votre système de gestion de la relation client (CRM) vers Salesforce.com.
- Réhéberger (lift and shift) : transférez une application vers le cloud sans apporter de modifications pour tirer parti des fonctionnalités du cloud. Exemple : migrez votre base de données Oracle sur site vers Oracle sur une instance EC2 dans le. AWS Cloud
- Relocaliser (lift and shift au niveau de l'hyperviseur) : transférez l'infrastructure vers le cloud sans acheter de nouveau matériel, réécrire des applications ou modifier vos opérations existantes. Vous migrez des serveurs d'une plateforme sur site vers un service cloud pour la même plateforme. Exemple : migrer une Microsoft Hyper-V application vers AWS.
- Retenir : conservez les applications dans votre environnement source. Il peut s'agir d'applications nécessitant une refactorisation majeure, que vous souhaitez retarder, et d'applications existantes que vous souhaitez retenir, car rien ne justifie leur migration sur le plan commercial.

- Retirer : mettez hors service ou supprimez les applications dont vous n'avez plus besoin dans votre environnement source.

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

Méthode de migration de base de données dans laquelle la synchronisation des bases de données source et cible est maintenue (à l'aide d'un outil de réplication bidirectionnelle ou d'opérations d'écriture double), tandis que les deux bases de données gèrent les transactions provenant de la connexion d'applications pendant la migration. Cette méthode prend en charge la migration par petits lots contrôlés au lieu d'exiger un basculement ponctuel. Elle est plus flexible mais demande plus de travail qu'une migration active-passive.

migration active-passive

Méthode de migration de base de données dans laquelle les bases de données source et cible sont synchronisées, mais seule la base de données source gère les transactions liées à la connexion des applications pendant que les données sont répliquées vers la base de données cible. La base de données cible n'accepte aucune transaction pendant la migration.

fonction d'agrégation

Fonction SQL qui agit sur un groupe de lignes et calcule une valeur de retour unique pour le groupe. Des exemples de fonctions d'agrégation incluent `SUM` et`MAX`.

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

Processus de suppression définitive d'informations personnelles dans un ensemble de données. L'anonymisation peut contribuer à protéger la vie privée. Les données anonymisées ne sont plus considérées comme des données personnelles.

anti-motif

Solution fréquemment utilisée pour un problème récurrent lorsque la solution est contre-productive, inefficace ou moins efficace qu'une alternative.

contrôle des applications

Une approche de sécurité qui permet d'utiliser uniquement des applications approuvées afin de protéger un système contre les logiciels malveillants.

portefeuille d'applications

Ensemble d'informations détaillées sur chaque application utilisée par une organisation, y compris le coût de génération et de maintenance de l'application, ainsi que sa valeur métier. Ces informations sont essentielles pour le processus de découverte et d'analyse du portefeuille et permettent d'identifier et de prioriser les applications à migrer, à moderniser et à optimiser.

intelligence artificielle (IA)

Domaine de l'informatique consacré à l'utilisation des technologies de calcul pour exécuter des fonctions cognitives généralement associées aux humains, telles que l'apprentissage, la résolution de problèmes et la reconnaissance de modèles. Pour plus d'informations, veuillez consulter Qu'est-ce que l'intelligence artificielle ?

opérations d'intelligence artificielle (AIOps)

Processus consistant à utiliser des techniques de machine learning pour résoudre les problèmes opérationnels, réduire les incidents opérationnels et les interventions humaines, mais aussi améliorer la qualité du service. Pour plus d'informations sur son AIOps utilisation dans la stratégie de AWS migration, consultez le guide d'intégration des opérations.

chiffrement asymétrique

Algorithme de chiffrement qui utilise une paire de clés, une clé publique pour le chiffrement et une clé privée pour le déchiffrement. Vous pouvez partager la clé publique, car elle n'est pas utilisée pour le déchiffrement, mais l'accès à la clé privée doit être très restreint.

atomicité, cohérence, isolement, durabilité (ACID)

Ensemble de propriétés logicielles garantissant la validité des données et la fiabilité opérationnelle d'une base de données, même en cas d'erreur, de panne de courant ou d'autres problèmes.

contrôle d'accès par attributs (ABAC)

Pratique qui consiste à créer des autorisations détaillées en fonction des attributs de l'utilisateur, tels que le service, le poste et le nom de l'équipe. Pour plus d'informations, consultez ABAC pour AWS dans la documentation Gestion des identités et des accès AWS (IAM).

source de données faisant autorité

Emplacement où vous stockez la version principale des données, considérée comme la source d'information la plus fiable. Vous pouvez copier les données de la source de données officielle vers d'autres emplacements à des fins de traitement ou de modification des données, par exemple en les anonymisant, en les expurgant ou en les pseudonymisant.

Zone de disponibilité

Un emplacement distinct au sein d'un Région AWS réseau isolé des défaillances dans d'autres zones de disponibilité et fournissant une connectivité réseau peu coûteuse et à faible latence aux autres zones de disponibilité de la même région.

AWS Cadre d'adoption du cloud (AWS CAF)

Un cadre de directives et de meilleures pratiques visant AWS à aider les entreprises à élaborer un plan efficace pour réussir leur migration vers le cloud. AWS La CAF organise ses conseils en six domaines prioritaires appelés perspectives : les affaires, les personnes, la gouvernance, les plateformes, la sécurité et les opérations. Les perspectives d'entreprise, de personnes et de gouvernance mettent l'accent sur les compétences et les processus métier, tandis que les perspectives relatives à la plateforme, à la sécurité et aux opérations se concentrent sur les compétences et les processus techniques. Par exemple, la perspective liée aux personnes cible les parties prenantes qui s'occupent des ressources humaines (RH), des fonctions de dotation en personnel et de la gestion des personnes. Dans cette perspective, la AWS CAF fournit des conseils pour le développement du personnel, la formation et les communications afin de préparer l'organisation à une adoption réussie du cloud. Pour plus d'informations, veuillez consulter le site Web AWS CAF et le livre blanc AWS CAF.

AWS Cadre de qualification de la charge de travail (AWS WQF)

Outil qui évalue les charges de travail liées à la migration des bases de données, recommande des stratégies de migration et fournit des estimations de travail. AWS Le WQF est inclus avec

AWS Schema Conversion Tool (AWS SCT). Il analyse les schémas de base de données et les objets de code, le code d'application, les dépendances et les caractéristiques de performance, et fournit des rapports d'évaluation.

B

mauvais bot

Un bot destiné à perturber ou à nuire à des individus ou à des organisations.

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

Vue unifiée et interactive des comportements des ressources et des interactions au fil du temps. Vous pouvez utiliser un graphique de comportement avec Amazon Detective pour examiner les tentatives de connexion infructueuses, les appels d'API suspects et les actions similaires. Pour plus d'informations, veuillez consulter Data in a behavior graph dans la documentation Detective.

système de poids fort

Système qui stocke d'abord l'octet le plus significatif. Voir aussi endianité.

classification binaire

Processus qui prédit un résultat binaire (l'une des deux classes possibles). Par exemple, votre modèle de machine learning peut avoir besoin de prévoir des problèmes tels que « Cet e-mail est-il du spam ou non ? » ou « Ce produit est-il un livre ou une voiture ? ».

filtre de Bloom

Structure de données probabiliste et efficace en termes de mémoire qui est utilisée pour tester si un élément fait partie d'un ensemble.

déploiement bleu/vert

Stratégie de déploiement dans laquelle vous créez deux environnements distincts mais identiques. Vous exécutez la version actuelle de l'application dans un environnement (bleu) et la nouvelle version de l'application dans l'autre environnement (vert). Cette stratégie vous permet de revenir rapidement en arrière avec un impact minimal.

bot

Application logicielle qui exécute des tâches automatisées sur Internet et simule l'activité ou l'interaction humaine. Certains robots sont utiles ou bénéfiques, comme les robots d'exploration Web qui indexent des informations sur Internet. D'autres robots, appelés « bots malveillants », sont destinés à perturber ou à nuire à des individus ou à des organisations.

botnet

Réseaux de robots infectés par des logiciels malveillants et contrôlés par une seule entité, connue sous le nom d'herder ou d'opérateur de bots. Les botnets sont le mécanisme le plus connu pour faire évoluer les bots et leur impact.

branche

Zone contenue d'un référentiel de code. La première branche créée dans un référentiel est la branche principale. Vous pouvez créer une branche à partir d'une branche existante, puis développer des fonctionnalités ou corriger des bogues dans la nouvelle branche. Une branche que vous créez pour générer une fonctionnalité est communément appelée branche de fonctionnalités. Lorsque la fonctionnalité est prête à être publiée, vous fusionnez à nouveau la branche de fonctionnalités dans la branche principale. Pour plus d'informations, consultez À propos des branches (GitHub documentation).

accès par brise-vitre

Dans des circonstances exceptionnelles et par le biais d'un processus approuvé, c'est un moyen rapide pour un utilisateur d'accéder à un accès auquel Compte AWS il n'est généralement pas autorisé. Pour plus d'informations, consultez l'indicateur Implementation break-glass procedures dans le guide Well-Architected AWS .

stratégie existante (brownfield)

L'infrastructure existante de votre environnement. Lorsque vous adoptez une stratégie existante pour une architecture système, vous concevez l'architecture en fonction des contraintes des systèmes et de l'infrastructure actuels. Si vous étendez l'infrastructure existante, vous pouvez combiner des politiques brownfield (existantes) et greenfield (inédites).

cache de tampon

Zone de mémoire dans laquelle sont stockées les données les plus fréquemment consultées.

capacité métier

Ce que fait une entreprise pour générer de la valeur (par exemple, les ventes, le service client ou le marketing). Les architectures de microservices et les décisions de développement peuvent être dictées par les capacités métier. Pour plus d'informations, veuillez consulter la section Organisation en fonction des capacités métier du livre blanc Exécution de microservices conteneurisés sur AWS.

planification de la continuité des activités (BCP)

Plan qui tient compte de l'impact potentiel d'un événement perturbateur, tel qu'une migration à grande échelle, sur les opérations, et qui permet à une entreprise de reprendre ses activités rapidement.

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

Diffusion lente et progressive d'une version pour les utilisateurs finaux. Lorsque vous êtes sûr, vous déployez la nouvelle version et remplacez la version actuelle dans son intégralité.

CCo E

Voir le Centre d'excellence du cloud.

CDC

Voir capture des données de modification.

capture des données de modification (CDC)

Processus de suivi des modifications apportées à une source de données, telle qu'une table de base de données, et d'enregistrement des métadonnées relatives à ces modifications. Vous pouvez utiliser la CDC à diverses fins, telles que l'audit ou la réplication des modifications dans un système cible afin de maintenir la synchronisation.

ingénierie du chaos

Introduire intentionnellement des défaillances ou des événements perturbateurs pour tester la résilience d'un système. Vous pouvez utiliser AWS Fault Injection Service (AWS FIS) pour effectuer des expériences qui stressent vos AWS charges de travail et évaluer leur réponse.

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

Processus de catégorisation qui permet de générer des prédictions. Les modèles de ML pour les problèmes de classification prédisent une valeur discrète. Les valeurs discrètes se distinguent toujours les unes des autres. Par exemple, un modèle peut avoir besoin d'évaluer la présence ou non d'une voiture sur une image.

chiffrement côté client

Chiffrement des données localement, avant que la cible ne les Service AWS reçoive.

Centre d'excellence du cloud (CCoE)

Une équipe multidisciplinaire qui dirige les efforts d'adoption du cloud au sein d'une organisation, notamment en développant les bonnes pratiques en matière de cloud, en mobilisant des ressources, en établissant des délais de migration et en guidant l'organisation dans le cadre de transformations à grande échelle. Pour plus d'informations, consultez les CCoarticles électroniques du blog sur la stratégie AWS Cloud d'entreprise.

cloud computing

Technologie cloud généralement utilisée pour le stockage de données à distance et la gestion des appareils IoT. Le cloud computing est généralement associé à la technologie informatique de pointe.

modèle d'exploitation du cloud

Dans une organisation informatique, modèle d'exploitation utilisé pour créer, faire évoluer et optimiser un ou plusieurs environnements cloud. Pour plus d'informations, consultez la section Création de votre modèle d'exploitation cloud.

étapes d'adoption du cloud

Les quatre phases que les entreprises traversent généralement lorsqu'elles migrent vers AWS Cloud :

- Projet : exécution de quelques projets liés au cloud à des fins de preuve de concept et d'apprentissage
- Base : réaliser des investissements fondamentaux pour accélérer votre adoption du cloud (par exemple, créer une zone de landing zone, définir un CCo E, établir un modèle opérationnel)
- Migration : migration d'applications individuelles
- Réinvention : optimisation des produits et services et innovation dans le cloud

Ces étapes ont été définies par Stephen Orban dans le billet de blog The Journey Toward Cloud-First & the Stages of Adoption publié sur le blog AWS Cloud Enterprise Strategy. Pour plus d'informations sur leur lien avec la stratégie de AWS migration, consultez le guide de préparation à la migration.

CMDB

Consultez la base de données de gestion des configurations.

référentiel de code

Emplacement où le code source et d'autres ressources, comme la documentation, les exemples et les scripts, sont stockés et mis à jour par le biais de processus de contrôle de version. Les référentiels cloud courants incluent GitHub ouBitbucket Cloud. Chaque version du code est appelée branche. Dans une structure de microservice, chaque référentiel est consacré à une seule fonctionnalité. Un seul pipeline CI/CD peut utiliser plusieurs référentiels.

cache passif

Cache tampon vide, mal rempli ou contenant des données obsolètes ou non pertinentes. Cela affecte les performances, car l'instance de base de données doit lire à partir de la mémoire principale ou du disque, ce qui est plus lent que la lecture à partir du cache tampon.

données gelées

Données rarement consultées et généralement historiques. Lorsque vous interrogez ce type de données, les requêtes lentes sont généralement acceptables. Le transfert de ces données vers des niveaux ou classes de stockage moins performants et moins coûteux peut réduire les coûts.

vision par ordinateur (CV)

Domaine de l'IA qui utilise l'apprentissage automatique pour analyser et extraire des informations à partir de formats visuels tels que des images numériques et des vidéos. Par exemple, Amazon SageMaker AI fournit des algorithmes de traitement d'image pour les CV.

dérive de configuration

Pour une charge de travail, une modification de configuration par rapport à l'état attendu. Cela peut entraîner une non-conformité de la charge de travail, et cela est généralement progressif et involontaire.

base de données de gestion des configurations (CMDB)

Référentiel qui stocke et gère les informations relatives à une base de données et à son environnement informatique, y compris les composants matériels et logiciels ainsi que leurs configurations. Vous utilisez généralement les données d'une CMDB lors de la phase de découverte et d'analyse du portefeuille de la migration.

pack de conformité

Ensemble de AWS Config règles et d'actions correctives que vous pouvez assembler pour personnaliser vos contrôles de conformité et de sécurité. Vous pouvez déployer un pack de conformité en tant qu'entité unique dans une région Compte AWS et, ou au sein d'une organisation, à l'aide d'un modèle YAML. Pour plus d'informations, consultez la section Packs de conformité dans la AWS Config documentation.

intégration continue et livraison continue (CI/CD)

Processus d'automatisation des étapes de source, de construction, de test, de préparation et de production du processus de publication du logiciel. CI/CD est communément décrit comme un pipeline. CI/CD peut vous aider à automatiser les processus, à améliorer la productivité, à améliorer la qualité du code et à accélérer les livraisons. Pour plus d'informations, veuillez consulter Avantages de la livraison continue. CD peut également signifier déploiement continu. Pour plus d'informations, veuillez consulter Livraison continue et déploiement continu.

CV

Voir vision par ordinateur.

D

données au repos

Données stationnaires dans votre réseau, telles que les données stockées.

classification des données

Processus permettant d'identifier et de catégoriser les données de votre réseau en fonction de leur sévérité et de leur sensibilité. Il s'agit d'un élément essentiel de toute stratégie de gestion des risques de cybersécurité, car il vous aide à déterminer les contrôles de protection et de conservation appropriés pour les données. La classification des données est une composante du pilier de sécurité du AWS Well-Architected Framework. Pour plus d'informations, veuillez consulter Classification des données.

dérive des données

Une variation significative entre les données de production et les données utilisées pour entraîner un modèle ML, ou une modification significative des données d'entrée au fil du temps. La dérive des données peut réduire la qualité, la précision et l'équité globales des prédictions des modèles ML.

données en transit

Données qui circulent activement sur votre réseau, par exemple entre les ressources du réseau.

maillage de données

Un cadre architectural qui fournit une propriété des données distribuée et décentralisée avec une gestion et une gouvernance centralisées.

minimisation des données

Le principe de collecte et de traitement des seules données strictement nécessaires. La pratique de la minimisation des données AWS Cloud peut réduire les risques liés à la confidentialité, les coûts et l'empreinte carbone de vos analyses.

périmètre de données

Ensemble de garde-fous préventifs dans votre AWS environnement qui permettent de garantir que seules les identités fiables accèdent aux ressources fiables des réseaux attendus. Pour plus d'informations, voir Création d'un périmètre de données sur AWS.

prétraitement des données

Pour transformer les données brutes en un format facile à analyser par votre modèle de ML. Le prétraitement des données peut impliquer la suppression de certaines colonnes ou lignes et le traitement des valeurs manquantes, incohérentes ou en double.

provenance des données

Le processus de suivi de l'origine et de l'historique des données tout au long de leur cycle de vie, par exemple la manière dont les données ont été générées, transmises et stockées.

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

Un système de gestion des données qui prend en charge les informations commerciales, telles que les analyses. Les entrepôts de données contiennent généralement de grandes quantités de données historiques et sont généralement utilisés pour les requêtes et les analyses.

langage de définition de base de données (DDL)

Instructions ou commandes permettant de créer ou de modifier la structure des tables et des objets dans une base de données.

langage de manipulation de base de données (DML)

Instructions ou commandes permettant de modifier (insérer, mettre à jour et supprimer) des informations dans une base de données.

DDL

Voir langage de définition de base de données.

ensemble profond

Sert à combiner plusieurs modèles de deep learning à des fins de prédiction. Vous pouvez utiliser des ensembles profonds pour obtenir une prévision plus précise ou pour estimer l'incertitude des prédictions.

deep learning

Un sous-champ de ML qui utilise plusieurs couches de réseaux neuronaux artificiels pour identifier le mappage entre les données d'entrée et les variables cibles d'intérêt.

defense-in-depth

Approche de la sécurité de l'information dans laquelle une série de mécanismes et de contrôles de sécurité sont judicieusement répartis sur l'ensemble d'un réseau informatique afin de protéger la confidentialité, l'intégrité et la disponibilité du réseau et des données qu'il contient. Lorsque vous adoptez cette stratégie AWS, vous ajoutez plusieurs contrôles à différentes couches de

la AWS Organizations structure afin de sécuriser les ressources. Par exemple, une defense-in-depth approche peut combiner l'authentification multifactorielle, la segmentation du réseau et le chiffrement.

administrateur délégué

Dans AWS Organizations, un service compatible peut enregistrer un compte AWS membre pour administrer les comptes de l'organisation et gérer les autorisations pour ce service. Ce compte est appelé administrateur délégué pour ce service. Pour plus d'informations et une liste des services compatibles, veuillez consulter la rubrique Services qui fonctionnent avec AWS Organizations dans la documentation AWS Organizations .

déploiement

Processus de mise à disposition d'une application, de nouvelles fonctionnalités ou de corrections de code dans l'environnement cible. Le déploiement implique la mise en œuvre de modifications dans une base de code, puis la génération et l'exécution de cette base de code dans les environnements de l'application.

environnement de développement

Voir environnement.

contrôle de détection

Contrôle de sécurité conçu pour détecter, journaliser et alerter après la survenue d'un événement. Ces contrôles constituent une deuxième ligne de défense et vous alertent en cas d'événements de sécurité qui ont contourné les contrôles préventifs en place. Pour plus d'informations, veuillez consulter la rubrique Contrôles de détection dans Implementing security controls on AWS.

cartographie de la chaîne de valeur du développement (DVSM)

Processus utilisé pour identifier et hiérarchiser les contraintes qui nuisent à la rapidité et à la qualité du cycle de vie du développement logiciel. DVSM étend le processus de cartographie de la chaîne de valeur initialement conçu pour les pratiques de production allégée. Il met l'accent sur les étapes et les équipes nécessaires pour créer et transférer de la valeur tout au long du processus de développement logiciel.

jumeau numérique

Représentation virtuelle d'un système réel, tel qu'un bâtiment, une usine, un équipement industriel ou une ligne de production. Les jumeaux numériques prennent en charge la maintenance prédictive, la surveillance à distance et l'optimisation de la production.

tableau des dimensions

Dans un schéma en étoile, table plus petite contenant les attributs de données relatifs aux données quantitatives d'une table de faits. Les attributs des tables de dimensions sont généralement des champs de texte ou des nombres discrets qui se comportent comme du texte. Ces attributs sont couramment utilisés pour la contrainte des requêtes, le filtrage et l'étiquetage des ensembles de résultats.

catastrophe

Un événement qui empêche une charge de travail ou un système d'atteindre ses objectifs commerciaux sur son site de déploiement principal. Ces événements peuvent être des catastrophes naturelles, des défaillances techniques ou le résultat d'actions humaines, telles qu'une mauvaise configuration involontaire ou une attaque de logiciel malveillant.

reprise après sinistre (DR)

La stratégie et le processus que vous utilisez pour minimiser les temps d'arrêt et les pertes de données causés par un sinistre. Pour plus d'informations, consultez Disaster Recovery of Workloads on AWS : Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

Approche visant à développer un système logiciel complexe en connectant ses composants à des domaines évolutifs, ou objectifs métier essentiels, que sert chaque composant. Ce concept a été introduit par Eric Evans dans son ouvrage Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston : Addison-Wesley Professional, 2003). Pour plus d'informations sur l'utilisation du design piloté par domaine avec le modèle de figuier étrangleur, veuillez consulter Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

DR

Voir reprise après sinistre.

détection de dérive

Suivi des écarts par rapport à une configuration de référence. Par exemple, vous pouvez l'utiliser AWS CloudFormation pour détecter la dérive des ressources du système ou AWS Control Tower

pour détecter les modifications de votre zone d'atterrissage susceptibles d'affecter le respect des exigences de gouvernance.

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

Technologie qui augmente la puissance de calcul des appareils intelligents en périphérie d'un réseau IoT. Comparé au cloud computing, l'informatique de pointe peut réduire la latence des communications et améliorer le temps de réponse.

échange de données informatisé (EDI)

L'échange automatique de documents commerciaux entre les organisations. Pour plus d'informations, voir Qu'est-ce que l'échange de données informatisé ?

chiffrement

Processus informatique qui transforme des données en texte clair, lisibles par l'homme, en texte chiffré.

clé de chiffrement

Chaîne cryptographique de bits aléatoires générée par un algorithme cryptographique. La longueur des clés peut varier, et chaque clé est conçue pour être imprévisible et unique.

endianisme

Ordre selon lequel les octets sont stockés dans la mémoire de l'ordinateur. Les systèmes de poids fort stockent d'abord l'octet le plus significatif. Les systèmes de poids faible stockent d'abord l'octet le moins significatif.

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

Service que vous pouvez héberger sur un cloud privé virtuel (VPC) pour le partager avec d'autres utilisateurs. Vous pouvez créer un service de point de terminaison avec AWS PrivateLink et accorder des autorisations à d'autres Comptes AWS ou à Gestion des identités et des accès AWS (IAM) principaux. Ces comptes ou principaux peuvent se connecter à votre service de point de terminaison de manière privée en créant des points de terminaison d'un VPC d'interface. Pour plus d'informations, veuillez consulter Création d'un service de point de terminaison dans la documentation Amazon Virtual Private Cloud (Amazon VPC).

planification des ressources d'entreprise (ERP)

Système qui automatise et gère les principaux processus métier (tels que la comptabilité, le MES et la gestion de projet) pour une entreprise.

chiffrement d'enveloppe

Processus de chiffrement d'une clé de chiffrement à l'aide d'une autre clé de chiffrement. Pour plus d'informations, consultez la section Chiffrement des enveloppes dans la documentation AWS Key Management Service (AWS KMS).

environment

Instance d'une application en cours d'exécution. Les types d'environnement les plus courants dans le cloud computing sont les suivants :

- Environnement de développement : instance d'une application en cours d'exécution à laquelle seule l'équipe principale chargée de la maintenance de l'application peut accéder. Les environnements de développement sont utilisés pour tester les modifications avant de les promouvoir dans les environnements supérieurs. Ce type d'environnement est parfois appelé environnement de test.
- Environnements inférieurs : tous les environnements de développement d'une application, tels que ceux utilisés pour les générations et les tests initiaux.
- Environnement de production : instance d'une application en cours d'exécution à laquelle les utilisateurs finaux peuvent accéder. Dans un CI/CD pipeline, l'environnement de production est le dernier environnement de déploiement.
- Environnements supérieurs : tous les environnements accessibles aux utilisateurs autres que l'équipe de développement principale. Ils peuvent inclure un environnement de production, des

environnements de préproduction et des environnements pour les tests d'acceptation par les utilisateurs.

époppée

Dans les méthodologies agiles, catégories fonctionnelles qui aident à organiser et à prioriser votre travail. Les épopées fournissent une description détaillée des exigences et des tâches d'implémentation. Par exemple, les points forts de la AWS CAF en matière de sécurité incluent la gestion des identités et des accès, les contrôles de détection, la sécurité des infrastructures, la protection des données et la réponse aux incidents. Pour plus d'informations sur les épopées dans la stratégie de migration AWS , veuillez consulter le guide d'implémentation du programme.

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

Processus d'analyse d'un jeu de données pour comprendre ses principales caractéristiques. Vous collectez ou agrégez des données, puis vous effectuez des enquêtes initiales pour trouver des modèles, détecter des anomalies et vérifier les hypothèses. L'EDA est réalisée en calculant des statistiques récapitulatives et en créant des visualisations de données.

F

tableau des faits

La table centrale dans un schéma en étoile. Il stocke des données quantitatives sur les opérations commerciales. Généralement, une table de faits contient deux types de colonnes : celles qui contiennent des mesures et celles qui contiennent une clé étrangère pour une table de dimensions.

échouer rapidement

Une philosophie qui utilise des tests fréquents et progressifs pour réduire le cycle de vie du développement. C'est un élément essentiel d'une approche agile.

limite d'isolation des défauts

Dans le AWS Cloud, une limite telle qu'une zone de disponibilité Région AWS, un plan de contrôle ou un plan de données qui limite l'effet d'une panne et contribue à améliorer la résilience des

charges de travail. Pour plus d'informations, consultez la section Limites d'isolation des AWS pannes.

branche de fonctionnalités

Voir succursale.

fonctionnalités

Les données d'entrée que vous utilisez pour faire une prédiction. Par exemple, dans un contexte de fabrication, les fonctionnalités peuvent être des images capturées périodiquement à partir de la ligne de fabrication.

importance des fonctionnalités

Le niveau d'importance d'une fonctionnalité pour les prédictions d'un modèle. Il s'exprime généralement sous la forme d'un score numérique qui peut être calculé à l'aide de différentes techniques, telles que la méthode Shapley Additive Explanations (SHAP) et les gradients intégrés. Pour plus d'informations, voir Interprétabilité du modèle d'apprentissage automatique avec AWS.

transformation de fonctionnalité

Optimiser les données pour le processus de ML, notamment en enrichissant les données avec des sources supplémentaires, en mettant à l'échelle les valeurs ou en extrayant plusieurs ensembles d'informations à partir d'un seul champ de données. Cela permet au modèle de ML de tirer parti des données. Par exemple, si vous décomposez la date « 2021-05-27 00:15:37 » en « 2021 », « mai », « jeudi » et « 15 », vous pouvez aider l'algorithme d'apprentissage à apprendre des modèles nuancés associés à différents composants de données.

invitation en quelques coups

Fournir à un LLM un petit nombre d'exemples illustrant la tâche et le résultat souhaité avant de lui demander d'effectuer une tâche similaire. Cette technique est une application de l'apprentissage contextuel, dans le cadre de laquelle les modèles apprennent à partir d'exemples (prises de vue) intégrés dans des instructions. Les instructions en quelques étapes peuvent être efficaces pour les tâches qui nécessitent un formatage, un raisonnement ou des connaissances de domaine spécifiques. Voir également l'invite Zero-Shot.

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

Utilisation de plusieurs conditions pour autoriser ou refuser une demande d'accès.

migration instantanée (flash-cut)

Méthode de migration de base de données qui utilise la réplication continue des données par le biais de la capture des données de modification afin de migrer les données dans les plus brefs délais, au lieu d'utiliser une approche progressive. L'objectif est de réduire au maximum les temps d'arrêt.

FM

Voir le modèle de fondation.

modèle de fondation (FM)

Un vaste réseau neuronal d'apprentissage profond qui s'est entraîné sur d'énormes ensembles de données généralisées et non étiquetées. FMs sont capables d'effectuer une grande variété de tâches générales, telles que comprendre le langage, générer du texte et des images et converser en langage naturel. Pour plus d'informations, voir Que sont les modèles de base ?

G

IA générative

Sous-ensemble de modèles d'IA qui ont été entraînés sur de grandes quantités de données et qui peuvent utiliser une simple invite textuelle pour créer de nouveaux contenus et artefacts, tels que des images, des vidéos, du texte et du son. Pour plus d'informations, consultez Qu'est-ce que l'IA générative.

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

Sur Amazon CloudFront, option permettant d'empêcher les utilisateurs de certains pays d'accéder aux distributions de contenu. Vous pouvez utiliser une liste d'autorisation ou une liste de blocage pour spécifier les pays approuvés et interdits. Pour plus d'informations, consultez la section Restreindre la distribution géographique de votre contenu dans la CloudFront documentation.

Flux de travail Gitflow

Approche dans laquelle les environnements inférieurs et supérieurs utilisent différentes branches dans un référentiel de code source. Le flux de travail Gitflow est considéré comme existant, et le flux de travail basé sur les troncs est l'approche moderne préférée.

image dorée

Un instantané d'un système ou d'un logiciel utilisé comme modèle pour déployer de nouvelles instances de ce système ou logiciel. Par exemple, dans le secteur de la fabrication, une image dorée peut être utilisée pour fournir des logiciels sur plusieurs appareils et contribue à améliorer la vitesse, l'évolutivité et la productivité des opérations de fabrication des appareils.

stratégie inédite

L'absence d'infrastructures existantes dans un nouvel environnement. Lorsque vous adoptez une stratégie inédite pour une architecture système, vous pouvez sélectionner toutes les nouvelles technologies sans restriction de compatibilité avec l'infrastructure existante, également appelée brownfield. Si vous étendez l'infrastructure existante, vous pouvez combiner des politiques brownfield (existantes) et greenfield (inédites).

barrière de protection

Règle de haut niveau qui permet de régir les ressources, les politiques et la conformité au sein des unités organisationnelles (OUs). Les barrières de protection préventives appliquent des politiques pour garantir l'alignement sur les normes de conformité. Elles sont mises en œuvre à l'aide de politiques de contrôle des services et de limites des autorisations IAM. Les barrières de protection de détection détectent les violations des politiques et les problèmes de conformité, et génèrent des alertes pour y remédier. Ils sont implémentés à l'aide d'Amazon AWS Config AWS Security Hub CSPM GuardDuty AWS Trusted Advisor, d'Amazon Inspector et de AWS Lambda contrôles personnalisés.

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

Migration de votre base de données source vers une base de données cible qui utilise un moteur de base de données différent (par exemple, Oracle vers Amazon Aurora). La migration hétérogène fait généralement partie d'un effort de réarchitecture, et la conversion du schéma peut s'avérer une tâche complexe. AWS propose AWS SCT qui facilite les conversions de schémas.

haute disponibilité (HA)

Capacité d'une charge de travail à fonctionner en continu, sans intervention, en cas de difficultés ou de catastrophes. Les systèmes HA sont conçus pour basculer automatiquement, fournir constamment des performances de haute qualité et gérer différentes charges et défaillances avec un impact minimal sur les performances.

modernisation des historiens

Approche utilisée pour moderniser et mettre à niveau les systèmes de technologie opérationnelle (OT) afin de mieux répondre aux besoins de l'industrie manufacturière. Un historien est un type de base de données utilisé pour collecter et stocker des données provenant de diverses sources dans une usine.

données de rétention

Partie de données historiques étiquetées qui n'est pas divulguée dans un ensemble de données utilisé pour entraîner un modèle d'apprentissage automatique. Vous pouvez utiliser les données de blocage pour évaluer les performances du modèle en comparant les prévisions du modèle aux données de blocage.

migration de base de données homogène

Migration de votre base de données source vers une base de données cible qui partage le même moteur de base de données (par exemple, Microsoft SQL Server vers Amazon RDS for SQL Server). La migration homogène s'inscrit généralement dans le cadre d'un effort de réhébergement ou de replateforme. Vous pouvez utiliser les utilitaires de base de données natifs pour migrer le schéma.

données chaudes

Données fréquemment consultées, telles que les données en temps réel ou les données translationnelles récentes. Ces données nécessitent généralement un niveau ou une classe de stockage à hautes performances pour fournir des réponses rapides aux requêtes.

correctif

Solution d'urgence à un problème critique dans un environnement de production. En raison de son urgence, un correctif est généralement créé en dehors du flux de travail de DevOps publication habituel.

période de soins intensifs

Immédiatement après le basculement, période pendant laquelle une équipe de migration gère et surveille les applications migrées dans le cloud afin de résoudre les problèmes éventuels. En règle générale, cette période dure de 1 à 4 jours. À la fin de la période de soins intensifs, l'équipe de migration transfère généralement la responsabilité des applications à l'équipe des opérations cloud.

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l'identité

Politique attachée à un ou plusieurs principaux IAM qui définit leurs autorisations au sein de l' AWS Cloud environnement.

application inactive

Application dont l'utilisation moyenne du processeur et de la mémoire se situe entre 5 et 20 % sur une période de 90 jours. Dans un projet de migration, il est courant de retirer ces applications ou de les retenir sur site.

IIo T

Voir Internet industriel des objets.

infrastructure immuable

Modèle qui déploie une nouvelle infrastructure pour les charges de travail de production au lieu de mettre à jour, d'appliquer des correctifs ou de modifier l'infrastructure existante. Les infrastructures immuables sont intrinsèquement plus cohérentes, fiables et prévisibles que les infrastructures mutables. Pour plus d'informations, consultez les meilleures pratiques de déploiement à l'aide d'une infrastructure immuable dans le AWS Well-Architected Framework.

VPC entrant (d'entrée)

Dans une architecture AWS multi-comptes, un VPC qui accepte, inspecte et achemine les connexions réseau depuis l'extérieur d'une application. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes

et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

migration incrémentielle

Stratégie de basculement dans le cadre de laquelle vous migrez votre application par petites parties au lieu d'effectuer un basculement complet unique. Par exemple, il se peut que vous ne transfériez que quelques microservices ou utilisateurs vers le nouveau système dans un premier temps. Après avoir vérifié que tout fonctionne correctement, vous pouvez transférer progressivement des microservices ou des utilisateurs supplémentaires jusqu'à ce que vous puissiez mettre hors service votre système hérité. Cette stratégie réduit les risques associés aux migrations de grande ampleur.

Industry 4.0

Terme introduit par Klaus Schwab en 2016 pour désigner la modernisation des processus de fabrication grâce aux avancées en matière de connectivité, de données en temps réel, d'automatisation, d'analyse et d'IA/ML.

infrastructure

Ensemble des ressources et des actifs contenus dans l'environnement d'une application.

infrastructure en tant que code (IaC)

Processus de mise en service et de gestion de l'infrastructure d'une application via un ensemble de fichiers de configuration. IaC est conçue pour vous aider à centraliser la gestion de l'infrastructure, à normaliser les ressources et à mettre à l'échelle rapidement afin que les nouveaux environnements soient reproductibles, fiables et cohérents.

Internet industriel des objets (IIoT)

L'utilisation de capteurs et d'appareils connectés à Internet dans les secteurs industriels tels que la fabrication, l'énergie, l'automobile, les soins de santé, les sciences de la vie et l'agriculture. Pour plus d'informations, voir Élaboration d'une stratégie de transformation numérique de l'Internet des objets (IIoT) industriel.

VPC d'inspection

Dans une architecture AWS multi-comptes, un VPC centralisé qui gère les inspections du trafic réseau VPCs entre (identique ou Régions AWS différent), Internet et les réseaux locaux. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau

avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

Internet des objets (IoT)

Réseau d'objets physiques connectés dotés de capteurs ou de processeurs intégrés qui communiquent avec d'autres appareils et systèmes via Internet ou via un réseau de communication local. Pour plus d'informations, veuillez consulter la section Qu'est-ce que l'IoT ?.

interprétabilité

Caractéristique d'un modèle de machine learning qui décrit dans quelle mesure un être humain peut comprendre comment les prédictions du modèle dépendent de ses entrées. Pour plus d'informations, voir Interprétabilité du modèle d'apprentissage automatique avec AWS.

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

Ensemble de bonnes pratiques pour proposer des services informatiques et les aligner sur les exigences métier. L'ITIL constitue la base de l'ITSM.

gestion des services informatiques (ITSM)

Activités associées à la conception, à la mise en œuvre, à la gestion et à la prise en charge de services informatiques d'une organisation. Pour plus d'informations sur l'intégration des opérations cloud aux outils ITSM, veuillez consulter le guide d'intégration des opérations.

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

Une implémentation du contrôle d'accès obligatoire (MAC) dans laquelle une valeur d'étiquette de sécurité est explicitement attribuée aux utilisateurs et aux données elles-mêmes. L'intersection

entre l'étiquette de sécurité utilisateur et l'étiquette de sécurité des données détermine les lignes et les colonnes visibles par l'utilisateur.

zone de destination

Une zone d'atterrissage est un AWS environnement multi-comptes bien conçu, évolutif et sécurisé. Il s'agit d'un point de départ à partir duquel vos entreprises peuvent rapidement lancer et déployer des charges de travail et des applications en toute confiance dans leur environnement de sécurité et d'infrastructure. Pour plus d'informations sur les zones de destination, veuillez consulter Setting up a secure and scalable multi-account AWS environment.

grand modèle de langage (LLM)

Un modèle d'intelligence artificielle basé sur le deep learning qui est préentraîné sur une grande quantité de données. Un LLM peut effectuer plusieurs tâches, telles que répondre à des questions, résumer des documents, traduire du texte dans d'autres langues et compléter des phrases. Pour plus d'informations, voir Que sont LLMs.

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

Bonne pratique de sécurité qui consiste à accorder les autorisations minimales nécessaires à l'exécution d'une tâche. Pour plus d'informations, veuillez consulter la rubrique Accorder les autorisations de moindre privilège dans la documentation IAM.

lift and shift

Voir 7 Rs.

système de poids faible

Système qui stocke d'abord l'octet le moins significatif. Voir aussi endianité.

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

Type d'intelligence artificielle qui utilise des algorithmes et des techniques pour la reconnaissance et l'apprentissage de modèles. Le ML analyse et apprend à partir de données enregistrées, telles que les données de l'Internet des objets (IoT), pour générer un modèle statistique basé sur des modèles. Pour plus d'informations, veuillez consulter Machine Learning.

branche principale

Voir succursale.

malware

Logiciel conçu pour compromettre la sécurité ou la confidentialité de l'ordinateur. Les logiciels malveillants peuvent perturber les systèmes informatiques, divulguer des informations sensibles ou obtenir un accès non autorisé. Parmi les malwares, on peut citer les virus, les vers, les rançongiciels, les chevaux de Troie, les logiciels espions et les enregistreurs de frappe.

services gérés

Services AWS pour lequel AWS fonctionnent la couche d'infrastructure, le système d'exploitation et les plateformes, et vous accédez aux points de terminaison pour stocker et récupérer des données. Amazon Simple Storage Service (Amazon S3) et Amazon DynamoDB sont des exemples de services gérés. Ils sont également connus sous le nom de services abstraits.

système d'exécution de la fabrication (MES)

Un système logiciel pour le suivi, la surveillance, la documentation et le contrôle des processus de production qui convertissent les matières premières en produits finis dans l'atelier.

MAP

Voir Migration Acceleration Program.

mécanisme

Processus complet au cours duquel vous créez un outil, favorisez son adoption, puis inspectez les résultats afin de procéder aux ajustements nécessaires. Un mécanisme est un cycle qui se renforce et s'améliore lorsqu'il fonctionne. Pour plus d'informations, voir Création de mécanismes dans le cadre AWS Well-Architected.

compte membre

Tous, à l' Comptes AWS exception du compte de gestion, qui font partie d'une organisation dans AWS Organizations. Un compte ne peut être membre que d'une seule organisation à la fois.

MAILLES

Voir le système d'exécution de la fabrication.

Transport télémétrique en file d'attente de messages (MQTT)

Protocole de communication léger machine-to-machine (M2M), basé sur le modèle de publication/d'abonnement, pour les appareils IoT aux ressources limitées.

microservice

Un petit service indépendant qui communique via un réseau bien défini APIs et qui est généralement détenu par de petites équipes autonomes. Par exemple, un système d'assurance peut inclure des microservices qui mappent à des capacités métier, telles que les ventes ou le marketing, ou à des sous-domaines, tels que les achats, les réclamations ou l'analytique. Les avantages des microservices incluent l'agilité, la flexibilité de la mise à l'échelle, la facilité de déploiement, la réutilisation du code et la résilience. Pour plus d'informations, consultez la section Intégration de microservices à l'aide de services AWS sans serveur.

architecture de microservices

Approche de création d'une application avec des composants indépendants qui exécutent chaque processus d'application en tant que microservice. Ces microservices communiquent via une interface bien définie en utilisant Lightweight. APIs Chaque microservice de cette architecture peut être mis à jour, déployé et mis à l'échelle pour répondre à la demande de fonctions spécifiques d'une application. Pour plus d'informations, consultez la section Implémentation de microservices sur AWS.

Programme d'accélération des migrations (MAP)

Un AWS programme qui fournit un support de conseil, des formations et des services pour aider les entreprises à établir une base opérationnelle solide pour passer au cloud, et pour aider à compenser le coût initial des migrations. MAP inclut une méthodologie de migration pour exécuter les migrations héritées de manière méthodique, ainsi qu'un ensemble d'outils pour automatiser et accélérer les scénarios de migration courants.

migration à grande échelle

Processus consistant à transférer la majeure partie du portefeuille d'applications vers le cloud par vagues, un plus grand nombre d'applications étant déplacées plus rapidement à chaque vague. Cette phase utilise les bonnes pratiques et les enseignements tirés des phases précédentes pour implémenter une usine de migration d'équipes, d'outils et de processus en vue de rationaliser la migration des charges de travail grâce à l'automatisation et à la livraison agile. Il s'agit de la troisième phase de la stratégie de migration AWS.

usine de migration

Équipes interfonctionnelles qui rationalisent la migration des charges de travail grâce à des approches automatisées et agiles. Les équipes de Migration Factory comprennent généralement des responsables des opérations, des analystes commerciaux et des propriétaires, des ingénieurs de migration, des développeurs et DevOps des professionnels travaillant dans le cadre de sprints. Entre 20 et 50 % du portefeuille d'applications d'entreprise est constitué de modèles répétés qui peuvent être optimisés par une approche d'usine. Pour plus d'informations, veuillez consulter la rubrique discussion of migration factories et le guide Cloud Migration Factory dans cet ensemble de contenus.

métadonnées de migration

Informations relatives à l'application et au serveur nécessaires pour finaliser la migration. Chaque modèle de migration nécessite un ensemble de métadonnées de migration différent. Les exemples de métadonnées de migration incluent le sous-réseau cible, le groupe de sécurité et le AWS compte.

modèle de migration

Tâche de migration reproductible qui détaille la stratégie de migration, la destination de la migration et l'application ou le service de migration utilisé. Exemple : réorganisez la migration vers Amazon EC2 AWS avec le service de migration d'applications.

Évaluation du portefeuille de migration (MPA)

Outil en ligne qui fournit des informations pour valider l'analyse de rentabilisation en faveur de la migration vers le. AWS Cloud La MPA propose une évaluation détaillée du portefeuille (dimensionnement approprié des serveurs, tarification, comparaison du coût total de possession, analyse des coûts de migration), ainsi que la planification de la migration (analyse et collecte des données d'applications, regroupement des applications, priorisation des migrations et planification des vagues). L'outil MPA (connexion requise) est disponible gratuitement pour tous les AWS consultants et consultants APN Partner.

Évaluation de la préparation à la migration (MRA)

Processus qui consiste à obtenir des informations sur l'état de préparation d'une organisation au cloud, à identifier les forces et les faiblesses et à élaborer un plan d'action pour combler les lacunes identifiées, à l'aide du AWS CAF. Pour plus d'informations, veuillez consulter le guide de préparation à la migration. La MRA est la première phase de la stratégie de migration AWS.

stratégie de migration

L'approche utilisée pour migrer une charge de travail vers le AWS Cloud. Pour plus d'informations, reportez-vous aux 7 R de ce glossaire et à Mobiliser votre organisation pour accélérer les migrations à grande échelle.

ML

Voir apprentissage automatique.

modernisation

Transformation d'une application obsolète (héritée ou monolithique) et de son infrastructure en un système agile, élastique et hautement disponible dans le cloud afin de réduire les coûts, de gagner en efficacité et de tirer parti des innovations. Pour plus d'informations, consultez la section Stratégie de modernisation des applications dans le AWS Cloud.

évaluation de la préparation à la modernisation

Évaluation qui permet de déterminer si les applications d'une organisation sont prêtes à être modernisées, d'identifier les avantages, les risques et les dépendances, et qui détermine dans quelle mesure l'organisation peut prendre en charge l'état futur de ces applications. Le résultat de l'évaluation est un plan de l'architecture cible, une feuille de route détaillant les phases de développement et les étapes du processus de modernisation, ainsi qu'un plan d'action pour combler les lacunes identifiées. Pour plus d'informations, consultez la section Évaluation de l'état de préparation à la modernisation des applications dans le AWS Cloud.

applications monolithiques (monolithes)

Applications qui s'exécutent en tant que service unique avec des processus étroitement couplés. Les applications monolithiques ont plusieurs inconvénients. Si une fonctionnalité de l'application connaît un pic de demande, l'architecture entière doit être mise à l'échelle. L'ajout ou l'amélioration des fonctionnalités d'une application monolithique devient également plus complexe lorsque la base de code s'élargit. Pour résoudre ces problèmes, vous pouvez utiliser une architecture de microservices. Pour plus d'informations, veuillez consulter Decomposing monoliths into microservices.

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

Processus qui permet de générer des prédictions pour plusieurs classes (prédiction d'un résultat parmi plus de deux). Par exemple, un modèle de ML peut demander « Ce produit est-il un livre, une voiture ou un téléphone ? » ou « Quelle catégorie de produits intéresse le plus ce client ? ».

infrastructure mutable

Modèle qui met à jour et modifie l'infrastructure existante pour les charges de travail de production. Pour améliorer la cohérence, la fiabilité et la prévisibilité, le AWS Well-Architected Framework recommande l'utilisation d'une infrastructure immuable comme meilleure pratique.

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

Méthode de migration dans laquelle la charge de travail source est supprimée au cours du processus de migration. Cette méthode implique un temps d'arrêt prolongé et est généralement utilisée pour de petites charges de travail non critiques.

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

Méthode de migration dans laquelle la charge de travail source est copiée sur le système cible sans être mise hors ligne. Les applications connectées à la charge de travail peuvent continuer à fonctionner pendant la migration. Cette méthode implique un temps d'arrêt nul ou minimal et est généralement utilisée pour les charges de travail de production critiques.

OPC-UA

Voir Open Process Communications - Architecture unifiée.

Communications par processus ouvert - Architecture unifiée (OPC-UA)

Un protocole de communication machine-to-machine (M2M) pour l'automatisation industrielle. L'OPC-UA fournit une norme d'interopérabilité avec des schémas de cryptage, d'authentification et d'autorisation des données.

accord au niveau opérationnel (OLA)

Accord qui précise ce que les groupes informatiques fonctionnels s'engagent à fournir les uns aux autres, afin de prendre en charge un contrat de niveau de service (SLA).

examen de l'état de préparation opérationnelle (ORR)

Une liste de questions et de bonnes pratiques associées qui vous aident à comprendre, à évaluer, à prévenir ou à réduire l'ampleur des incidents et des défaillances possibles. Pour plus d'informations, voir Operational Readiness Reviews (ORR) dans le AWS Well-Architected Framework.

technologie opérationnelle (OT)

Systèmes matériels et logiciels qui fonctionnent avec l'environnement physique pour contrôler les opérations, les équipements et les infrastructures industriels. Dans le secteur manufacturier, l'intégration des systèmes OT et des technologies de l'information (IT) est au cœur des transformations de l'industrie 4.0.

intégration des opérations (OI)

Processus de modernisation des opérations dans le cloud, qui implique la planification de la préparation, l'automatisation et l'intégration. Pour en savoir plus, veuillez consulter le guide d'intégration des opérations.

journal de suivi d'organisation

Un parcours créé par AWS CloudTrail qui enregistre tous les événements pour tous les membres Comptes AWS d'une organisation dans AWS Organizations. Ce journal de suivi est créé dans chaque Compte AWS qui fait partie de l'organisation et suit l'activité de chaque compte. Pour plus d'informations, consultez la section Création d'un suivi pour une organisation dans la CloudTrail documentation.

gestion du changement organisationnel (OCM)

Cadre pour gérer les transformations métier majeures et perturbatrices du point de vue des personnes, de la culture et du leadership. L'OCM aide les organisations à se préparer et à effectuer la transition vers de nouveaux systèmes et de nouvelles politiques en accélérant l'adoption des changements, en abordant les problèmes de transition et en favorisant des changements culturels et organisationnels. Dans la stratégie de AWS migration, ce cadre est appelé accélération du personnel, en raison de la rapidité du changement requise dans les projets d'adoption du cloud. Pour plus d'informations, veuillez consulter le guide OCM.

contrôle d'accès d'origine (OAC)

Dans CloudFront, une option améliorée pour restreindre l'accès afin de sécuriser votre contenu Amazon Simple Storage Service (Amazon S3). L'OAC prend en charge tous les compartiments S3 dans leur ensemble Régions AWS, le chiffrement côté serveur avec AWS KMS (SSE-KMS) et les requêtes dynamiques `PUT` adressées au compartiment S3. `DELETE`

identité d'accès d'origine (OAI)

Dans CloudFront, une option permettant de restreindre l'accès afin de sécuriser votre contenu Amazon S3. Lorsque vous utilisez OAI, il CloudFront crée un principal auprès duquel Amazon S3 peut s'authentifier. Les principaux authentifiés peuvent accéder au contenu d'un compartiment S3 uniquement via une distribution spécifique CloudFront . Voir également OAC, qui fournit un contrôle d'accès plus précis et amélioré.

ORR

Voir l'examen de l'état de préparation opérationnelle.

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

Dans une architecture AWS multi-comptes, un VPC qui gère les connexions réseau initiées depuis une application. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

P

limite des autorisations

Politique de gestion IAM attachée aux principaux IAM pour définir les autorisations maximales que peut avoir l'utilisateur ou le rôle. Pour plus d'informations, veuillez consulter la rubrique Limites des autorisations dans la documentation IAM.

informations personnelles identifiables (PII)

Informations qui, lorsqu'elles sont consultées directement ou associées à d'autres données connexes, peuvent être utilisées pour déduire raisonnablement l'identité d'une personne. Les exemples d'informations personnelles incluent les noms, les adresses et les informations de contact.

PII

Voir les informations personnelles identifiables.

manuel stratégique

Ensemble d'étapes prédéfinies qui capturent le travail associé aux migrations, comme la fourniture de fonctions d'opérations de base dans le cloud. Un manuel stratégique peut revêtir la forme de scripts, de runbooks automatisés ou d'un résumé des processus ou des étapes nécessaires au fonctionnement de votre environnement modernisé.

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

policy

Objet capable de définir les autorisations (voir la politique basée sur l'identité), de spécifier les conditions d'accès (voir la politique basée sur les ressources) ou de définir les autorisations maximales pour tous les comptes d'une organisation dans AWS Organizations (voir la politique de contrôle des services).

persistance polyglotte

Choix indépendant de la technologie de stockage de données d'un microservice en fonction des modèles d'accès aux données et d'autres exigences. Si vos microservices utilisent la même technologie de stockage de données, ils peuvent rencontrer des difficultés d'implémentation ou présenter des performances médiocres. Les microservices sont plus faciles à mettre en œuvre, atteignent de meilleures performances, ainsi qu'une meilleure capacité de mise à l'échelle s'ils utilisent l'entrepôt de données le mieux adapté à leurs besoins.

évaluation du portefeuille

Processus de découverte, d'analyse et de priorisation du portefeuille d'applications afin de planifier la migration. Pour plus d'informations, veuillez consulter Evaluating migration readiness.

predicate

Une condition de requête qui renvoie `true` ou`false`, généralement située dans une `WHERE` clause.

prédicat pushdown

Technique d'optimisation des requêtes de base de données qui filtre les données de la requête avant le transfert. Cela réduit la quantité de données qui doivent être extraites et traitées à partir de la base de données relationnelle et améliore les performances des requêtes.

contrôle préventif

Contrôle de sécurité conçu pour empêcher qu'un événement ne se produise. Ces contrôles constituent une première ligne de défense pour empêcher tout accès non autorisé ou toute modification indésirable de votre réseau. Pour plus d'informations, veuillez consulter Preventative controls dans Implementing security controls on AWS.

principal

Entité AWS capable d'effectuer des actions et d'accéder aux ressources. Cette entité est généralement un utilisateur root pour un Compte AWS rôle IAM ou un utilisateur. Pour plus

d'informations, veuillez consulter la rubrique Principal dans Termes et concepts relatifs aux rôles, dans la documentation IAM.

confidentialité dès la conception

Une approche d'ingénierie système qui prend en compte la confidentialité tout au long du processus de développement.

zones hébergées privées

Conteneur contenant des informations sur la manière dont vous souhaitez qu'Amazon Route 53 réponde aux requêtes DNS pour un domaine et ses sous-domaines au sein d'un ou de plusieurs VPCs domaines. Pour plus d'informations, veuillez consulter Working with private hosted zones dans la documentation Route 53.

contrôle proactif

Contrôle de sécurité conçu pour empêcher le déploiement de ressources non conformes. Ces contrôles analysent les ressources avant qu'elles ne soient provisionnées. Si la ressource n'est pas conforme au contrôle, elle n'est pas provisionnée. Pour plus d'informations, consultez le guide de référence sur les contrôles dans la AWS Control Tower documentation et consultez la section Contrôles proactifs dans Implémentation des contrôles de sécurité sur AWS.

gestion du cycle de vie des produits (PLM)

Gestion des données et des processus d'un produit tout au long de son cycle de vie, depuis la conception, le développement et le lancement, en passant par la croissance et la maturité, jusqu'au déclin et au retrait.

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

Dans le secteur manufacturier, un ordinateur hautement fiable et adaptable qui surveille les machines et automatise les processus de fabrication.

chaînage rapide

Utiliser le résultat d'une invite LLM comme entrée pour l'invite suivante afin de générer de meilleures réponses. Cette technique est utilisée pour décomposer une tâche complexe en sous-tâches ou pour affiner ou développer de manière itérative une réponse préliminaire. Cela permet d'améliorer la précision et la pertinence des réponses d'un modèle et permet d'obtenir des résultats plus précis et personnalisés.

pseudonymisation

Processus de remplacement des identifiants personnels dans un ensemble de données par des valeurs fictives. La pseudonymisation peut contribuer à protéger la vie privée. Les données pseudonymisées sont toujours considérées comme des données personnelles.

publish/subscribe (pub/sub)

Modèle qui permet des communications asynchrones entre les microservices afin d'améliorer l'évolutivité et la réactivité. Par exemple, dans un MES basé sur des microservices, un microservice peut publier des messages d'événements sur un canal auquel d'autres microservices peuvent s'abonner. Le système peut ajouter de nouveaux microservices sans modifier le service de publication.

Q

plan de requête

Série d'étapes, telles que des instructions, utilisées pour accéder aux données d'un système de base de données relationnelle SQL.

régression du plan de requêtes

Le cas où un optimiseur de service de base de données choisit un plan moins optimal qu'avant une modification donnée de l'environnement de base de données. Cela peut être dû à des changements en termes de statistiques, de contraintes, de paramètres d'environnement, de liaisons de paramètres de requêtes et de mises à jour du moteur de base de données.

R

Matrice RACI

Voir responsable, responsable, consulté, informé (RACI).

RAG

Voir Retrieval Augmented Generation.

rançongiciel

Logiciel malveillant conçu pour bloquer l'accès à un système informatique ou à des données jusqu'à ce qu'un paiement soit effectué.

Matrice RASCI

Voir responsable, responsable, consulté, informé (RACI).

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

Copie d'une base de données utilisée en lecture seule. Vous pouvez acheminer les requêtes vers le réplica de lecture pour réduire la charge sur votre base de données principale.

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

Durée maximale acceptable depuis le dernier point de récupération des données. Il détermine ce qui est considéré comme étant une perte de données acceptable entre le dernier point de reprise et l'interruption du service.

objectif de temps de récupération (RTO)

Le délai maximum acceptable entre l'interruption du service et le rétablissement du service.

refactoriser

Voir 7 Rs.

Région

Un ensemble de AWS ressources dans une zone géographique. Chacun Région AWS est isolé et indépendant des autres pour garantir tolérance aux pannes, stabilité et résilience. Pour plus d'informations, voir Spécifier ce que Régions AWS votre compte peut utiliser.

régression

Technique de ML qui prédit une valeur numérique. Par exemple, pour résoudre le problème « Quel sera le prix de vente de cette maison ? », un modèle de ML pourrait utiliser un modèle de régression linéaire pour prédire le prix de vente d'une maison sur la base de faits connus à son sujet (par exemple, la superficie en mètres carrés).

réhéberger

Voir 7 Rs.

version

Dans un processus de déploiement, action visant à promouvoir les modifications apportées à un environnement de production.

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

La capacité d'une application à résister aux perturbations ou à s'en remettre. La haute disponibilité et la reprise après sinistre sont des considérations courantes lors de la planification de la résilience dans le AWS Cloud. Pour plus d'informations, consultez AWS Cloud Résilience.

politique basée sur les ressources

Politique attachée à une ressource, comme un compartiment Amazon S3, un point de terminaison ou une clé de chiffrement. Ce type de politique précise les principaux auxquels l'accès est autorisé, les actions prises en charge et toutes les autres conditions qui doivent être remplies.

matrice responsable, redevable, consulté et informé (RACI)

Une matrice qui définit les rôles et les responsabilités de toutes les parties impliquées dans les activités de migration et les opérations cloud. Le nom de la matrice est dérivé des types de responsabilité définis dans la matrice : responsable (R), responsable (A), consulté (C) et informé (I). Le type de support (S) est facultatif. Si vous incluez le support, la matrice est appelée matrice RASCI, et si vous l'excluez, elle est appelée matrice RACI.

contrôle réactif

Contrôle de sécurité conçu pour permettre de remédier aux événements indésirables ou aux écarts par rapport à votre référence de sécurité. Pour plus d'informations, veuillez consulter la rubrique Responsive controls dans Implementing security controls on AWS.

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

Génération augmentée de récupération (RAG)

Technologie d'IA générative dans laquelle un LLM fait référence à une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation avant de générer une réponse. Par exemple, un modèle RAG peut effectuer une recherche sémantique dans la base de connaissances ou dans les données personnalisées d'une organisation. Pour plus d'informations, voir Qu'est-ce que RAG ?

rotation

Processus de mise à jour périodique d'un secret pour empêcher un attaquant d'accéder aux informations d'identification.

contrôle d'accès aux lignes et aux colonnes (RCAC)

Utilisation d'expressions SQL simples et flexibles dotées de règles d'accès définies. Le RCAC comprend des autorisations de ligne et des masques de colonnes.

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif en matière de temps de rétablissement.

runbook

Ensemble de procédures manuelles ou automatisées nécessaires à l'exécution d'une tâche spécifique. Elles visent généralement à rationaliser les opérations ou les procédures répétitives présentant des taux d'erreur élevés.

S

SAML 2.0

Un standard ouvert utilisé par de nombreux fournisseurs d'identité (IdPs). Cette fonctionnalité permet l'authentification unique fédérée (SSO), afin que les utilisateurs puissent se connecter

AWS Management Console ou appeler les opérations de l' AWS API sans que vous ayez à créer un utilisateur dans IAM pour tous les membres de votre organisation. Pour plus d'informations sur la fédération SAML 2.0, veuillez consulter À propos de la fédération SAML 2.0 dans la documentation IAM.

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

Dans AWS Secrets Manager des informations confidentielles ou restreintes, telles qu'un mot de passe ou des informations d'identification utilisateur, que vous stockez sous forme cryptée. Il comprend la valeur secrète et ses métadonnées. La valeur secrète peut être binaire, une chaîne unique ou plusieurs chaînes. Pour plus d'informations, voir Que contient le secret d'un Secrets Manager ? dans la documentation de Secrets Manager.

sécurité dès la conception

Une approche d'ingénierie système qui prend en compte la sécurité tout au long du processus de développement.

contrôle de sécurité

Barrière de protection technique ou administrative qui empêche, détecte ou réduit la capacité d'un assaillant d'exploiter une vulnérabilité de sécurité. Il existe quatre principaux types de contrôles de sécurité : préventifs, détectifs, réactifs et proactifs.

renforcement de la sécurité

Processus qui consiste à réduire la surface d'attaque pour la rendre plus résistante aux attaques. Cela peut inclure des actions telles que la suppression de ressources qui ne sont plus requises, la mise en œuvre des bonnes pratiques de sécurité consistant à accorder le moindre privilège ou la désactivation de fonctionnalités inutiles dans les fichiers de configuration.

système de gestion des informations et des événements de sécurité (SIEM)

Outils et services qui associent les systèmes de gestion des informations de sécurité (SIM) et de gestion des événements de sécurité (SEM). Un système SIEM collecte, surveille et analyse les

données provenant de serveurs, de réseaux, d'appareils et d'autres sources afin de détecter les menaces et les failles de sécurité, mais aussi de générer des alertes.

automatisation des réponses de sécurité

Action prédéfinie et programmée conçue pour répondre automatiquement à un événement de sécurité ou y remédier. Ces automatisations servent de contrôles de sécurité détectifs ou réactifs qui vous aident à mettre en œuvre les meilleures pratiques en matière AWS de sécurité. Parmi les actions de réponse automatique, citons la modification d'un groupe de sécurité VPC, l'application de correctifs à une instance Amazon EC2 ou la rotation des informations d'identification.

chiffrement côté serveur

Chiffrement des données à destination, par celui Service AWS qui les reçoit.

Politique de contrôle des services (SCP)

Politique qui fournit un contrôle centralisé des autorisations pour tous les comptes d'une organisation dans AWS Organizations. SCPs définissez des garde-fous ou des limites aux actions qu'un administrateur peut déléguer à des utilisateurs ou à des rôles. Vous pouvez les utiliser SCPs comme listes d'autorisation ou de refus pour spécifier les services ou les actions autorisés ou interdits. Pour plus d'informations, consultez la section Politiques de contrôle des services dans la AWS Organizations documentation.

point de terminaison du service

URL du point d'entrée pour un Service AWS. Pour vous connecter par programmation au service cible, vous pouvez utiliser un point de terminaison. Pour plus d'informations, veuillez consulter la rubrique Service AWS endpoints dans Références générales AWS.

contrat de niveau de service (SLA)

Accord qui précise ce qu'une équipe informatique promet de fournir à ses clients, comme le temps de disponibilité et les performances des services.

indicateur de niveau de service (SLI)

Mesure d'un aspect des performances d'un service, tel que son taux d'erreur, sa disponibilité ou son débit.

objectif de niveau de service (SLO)

Mesure cible qui représente l'état d'un service, tel que mesuré par un indicateur de niveau de service.

modèle de responsabilité partagée

Un modèle décrivant la responsabilité que vous partagez en matière AWS de sécurité et de conformité dans le cloud. AWS est responsable de la sécurité du cloud, alors que vous êtes responsable de la sécurité dans le cloud. Pour de plus amples informations, veuillez consulter Modèle de responsabilité partagée.

SIEM

Consultez les informations de sécurité et le système de gestion des événements.

point de défaillance unique (SPOF)

Défaillance d'un seul composant critique d'une application susceptible de perturber le système.

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

Modèle permettant de mettre à l'échelle et d'accélérer les projets de modernisation. Au fur et à mesure que les nouvelles fonctionnalités et les nouvelles versions de produits sont définies, l'équipe principale se divise pour créer des équipes de produit. Cela permet de mettre à l'échelle les capacités et les services de votre organisation, d'améliorer la productivité des développeurs et de favoriser une innovation rapide. Pour plus d'informations, voir Approche progressive de la modernisation des applications dans le. AWS Cloud

SPOF

Voir point de défaillance unique.

schéma en étoile

Structure organisationnelle de base de données qui utilise une grande table de faits pour stocker les données transactionnelles ou mesurées et utilise une ou plusieurs tables dimensionnelles plus petites pour stocker les attributs des données. Cette structure est conçue pour être utilisée dans un entrepôt de données ou à des fins de business intelligence.

modèle de figuier étrangleur

Approche de modernisation des systèmes monolithiques en réécrivant et en remplaçant progressivement les fonctionnalités du système jusqu'à ce que le système hérité puisse être mis hors service. Ce modèle utilise l'analogie d'un figuier de vigne qui se développe dans un arbre existant et qui finit par supplanter son hôte. Le schéma a été présenté par Martin Fowler comme un moyen de gérer les risques lors de la réécriture de systèmes monolithiques. Pour obtenir un exemple d'application de ce modèle, veuillez consulter Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

sous-réseau

Plage d'adresses IP dans votre VPC. Un sous-réseau doit se trouver dans une seule zone de disponibilité.

contrôle de supervision et acquisition de données (SCADA)

Dans le secteur manufacturier, un système qui utilise du matériel et des logiciels pour surveiller les actifs physiques et les opérations de production.

chiffrement symétrique

Algorithme de chiffrement qui utilise la même clé pour chiffrer et déchiffrer les données.

tests synthétiques

Tester un système de manière à simuler les interactions des utilisateurs afin de détecter les problèmes potentiels ou de surveiller les performances. Vous pouvez utiliser Amazon CloudWatch Synthetics pour créer ces tests.

invite du système

Technique permettant de fournir un contexte, des instructions ou des directives à un LLM afin d'orienter son comportement. Les instructions du système aident à définir le contexte et à établir des règles pour les interactions avec les utilisateurs.

T

tags

Des paires clé-valeur qui agissent comme des métadonnées pour organiser vos AWS ressources. Les balises peuvent vous aider à gérer, identifier, organiser, rechercher et filtrer des ressources. Pour plus d’informations, veuillez consulter la rubrique Balisage de vos AWS ressources.

variable cible

La valeur que vous essayez de prédire dans le cadre du ML supervisé. Elle est également qualifiée de variable de résultat. Par exemple, dans un environnement de fabrication, la variable cible peut être un défaut du produit.

liste de tâches

Outil utilisé pour suivre les progrès dans un runbook. Liste de tâches qui contient une vue d'ensemble du runbook et une liste des tâches générales à effectuer. Pour chaque tâche générale, elle inclut le temps estimé nécessaire, le propriétaire et l'avancement.

environnement de test

Voir environnement.

entraînement

Pour fournir des données à partir desquelles votre modèle de ML peut apprendre. Les données d'entraînement doivent contenir la bonne réponse. L'algorithme d'apprentissage identifie des modèles dans les données d'entraînement, qui mettent en correspondance les attributs des données d'entrée avec la cible (la réponse que vous souhaitez prédire). Il fournit un modèle de ML qui capture ces modèles. Vous pouvez alors utiliser le modèle de ML pour obtenir des prédictions sur de nouvelles données pour lesquelles vous ne connaissez pas la cible.

passerelle de transit

Un hub de transit réseau que vous pouvez utiliser pour interconnecter vos réseaux VPCs et ceux sur site. Pour plus d'informations, voir Qu'est-ce qu'une passerelle de transit dans la AWS Transit Gateway documentation.

flux de travail basé sur jonction

Approche selon laquelle les développeurs génèrent et testent des fonctionnalités localement dans une branche de fonctionnalités, puis fusionnent ces modifications dans la branche principale. La branche principale est ensuite intégrée aux environnements de développement, de préproduction et de production, de manière séquentielle.

accès sécurisé

Accorder des autorisations à un service que vous spécifiez pour effectuer des tâches au sein de votre organisation AWS Organizations et dans ses comptes en votre nom. Le service de confiance crée un rôle lié au service dans chaque compte, lorsque ce rôle est nécessaire, pour effectuer des tâches de gestion à votre place. Pour plus d'informations, consultez la

section Utilisation AWS Organizations avec d'autres AWS services dans la AWS Organizations documentation.

réglage

Pour modifier certains aspects de votre processus d'entraînement afin d'améliorer la précision du modèle de ML. Par exemple, vous pouvez entraîner le modèle de ML en générant un ensemble d'étiquetage, en ajoutant des étiquettes, puis en répétant ces étapes plusieurs fois avec différents paramètres pour optimiser le modèle.

équipe de deux pizzas

Une petite DevOps équipe que vous pouvez nourrir avec deux pizzas. Une équipe de deux pizzas garantit les meilleures opportunités de collaboration possible dans le développement de logiciels.

U

incertitude

Un concept qui fait référence à des informations imprécises, incomplètes ou inconnues susceptibles de compromettre la fiabilité des modèles de ML prédictifs. Il existe deux types d'incertitude : l'incertitude épistémique est causée par des données limitées et incomplètes, alors que l'incertitude aléatoire est causée par le bruit et le caractère aléatoire inhérents aux données.

tâches indifférenciées

Également connu sous le nom de « levage de charges lourdes », ce travail est nécessaire pour créer et exploiter une application, mais qui n'apporte pas de valeur directe à l'utilisateur final ni d'avantage concurrentiel. Les exemples de tâches indifférenciées incluent l'approvisionnement, la maintenance et la planification des capacités.

environnements supérieurs

Voir environnement.

V

mise à vide

Opération de maintenance de base de données qui implique un nettoyage après des mises à jour incrémentielles afin de récupérer de l'espace de stockage et d'améliorer les performances.

contrôle de version

Processus et outils permettant de suivre les modifications, telles que les modifications apportées au code source dans un référentiel.

Appairage de VPC

Une connexion entre deux VPCs qui vous permet d'acheminer le trafic en utilisant des adresses IP privées. Pour plus d'informations, veuillez consulter la rubrique Qu'est-ce que l'appairage de VPC ? dans la documentation Amazon VPC.

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

Cache tampon qui contient les données actuelles et pertinentes fréquemment consultées. L'instance de base de données peut lire à partir du cache tampon, ce qui est plus rapide que la lecture à partir de la mémoire principale ou du disque.

données chaudes

Données rarement consultées. Lorsque vous interrogez ce type de données, des requêtes modérément lentes sont généralement acceptables.

fonction de fenêtre

Fonction SQL qui effectue un calcul sur un groupe de lignes liées d'une manière ou d'une autre à l'enregistrement en cours. Les fonctions de fenêtre sont utiles pour traiter des tâches, telles que le calcul d'une moyenne mobile ou l'accès à la valeur des lignes en fonction de la position relative de la ligne en cours.

charge de travail

Ensemble de ressources et de code qui fournit une valeur métier, par exemple une application destinée au client ou un processus de backend.

flux de travail

Groupes fonctionnels d'un projet de migration chargés d'un ensemble de tâches spécifique. Chaque flux de travail est indépendant, mais prend en charge les autres flux de travail du projet.

Par exemple, le flux de travail du portefeuille est chargé de prioriser les applications, de planifier les vagues et de collecter les métadonnées de migration. Le flux de travail du portefeuille fournit ces actifs au flux de travail de migration, qui migre ensuite les serveurs et les applications.

VER

Voir écrire une fois, lire plusieurs.

WQF

Voir le cadre AWS de qualification de la charge de travail.

écrire une fois, lire plusieurs (WORM)

Modèle de stockage qui écrit les données une seule fois et empêche leur suppression ou leur modification. Les utilisateurs autorisés peuvent lire les données autant de fois que nécessaire, mais ils ne peuvent pas les modifier. Cette infrastructure de stockage de données est considérée comme immuable.

Z

exploit Zero-Day

Une attaque, généralement un logiciel malveillant, qui tire parti d'une vulnérabilité de type « jour zéro ».

vulnérabilité « jour zéro »

Une faille ou une vulnérabilité non atténuée dans un système de production. Les acteurs malveillants peuvent utiliser ce type de vulnérabilité pour attaquer le système. Les développeurs prennent souvent conscience de la vulnérabilité à la suite de l'attaque.

invite Zero-Shot

Fournir à un LLM des instructions pour effectuer une tâche, mais aucun exemple (plans) pouvant aider à la guider. Le LLM doit utiliser ses connaissances pré-entraînées pour gérer la tâche. L'efficacité de l'invite zéro dépend de la complexité de la tâche et de la qualité de l'invite. Voir également les instructions en quelques clics.

application zombie

Application dont l'utilisation moyenne du processeur et de la mémoire est inférieure à 5 %. Dans un projet de migration, il est courant de retirer ces applications.

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.