TechTarget | Custom Media
E-Book

Comment l'IA change tout, du centre de données au cloud en passant par la bordure de l'Internet

L'impact de l'IA en termes de transformation est immense. Elle modifie la nature même de la manière dont le cloud et l'Edge Computing apportent de la valeur aux entreprises, à travers tous les secteurs, toutes les régions et tous les cas d'utilisation. Ce livre numérique donne un point de vue pratique sur le rôle de plus en plus important de l'IA dans le cloud distribué.

Akamai

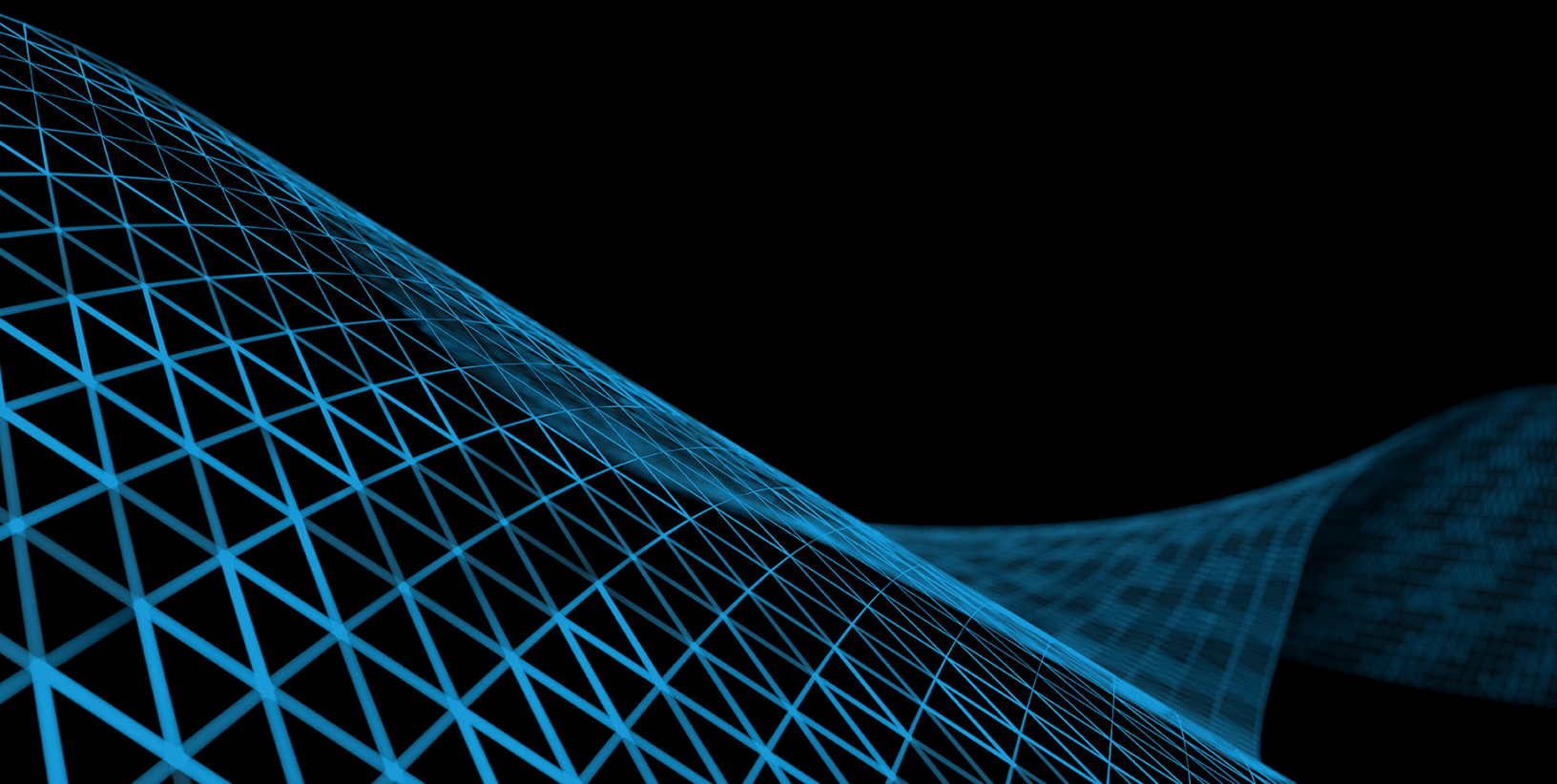
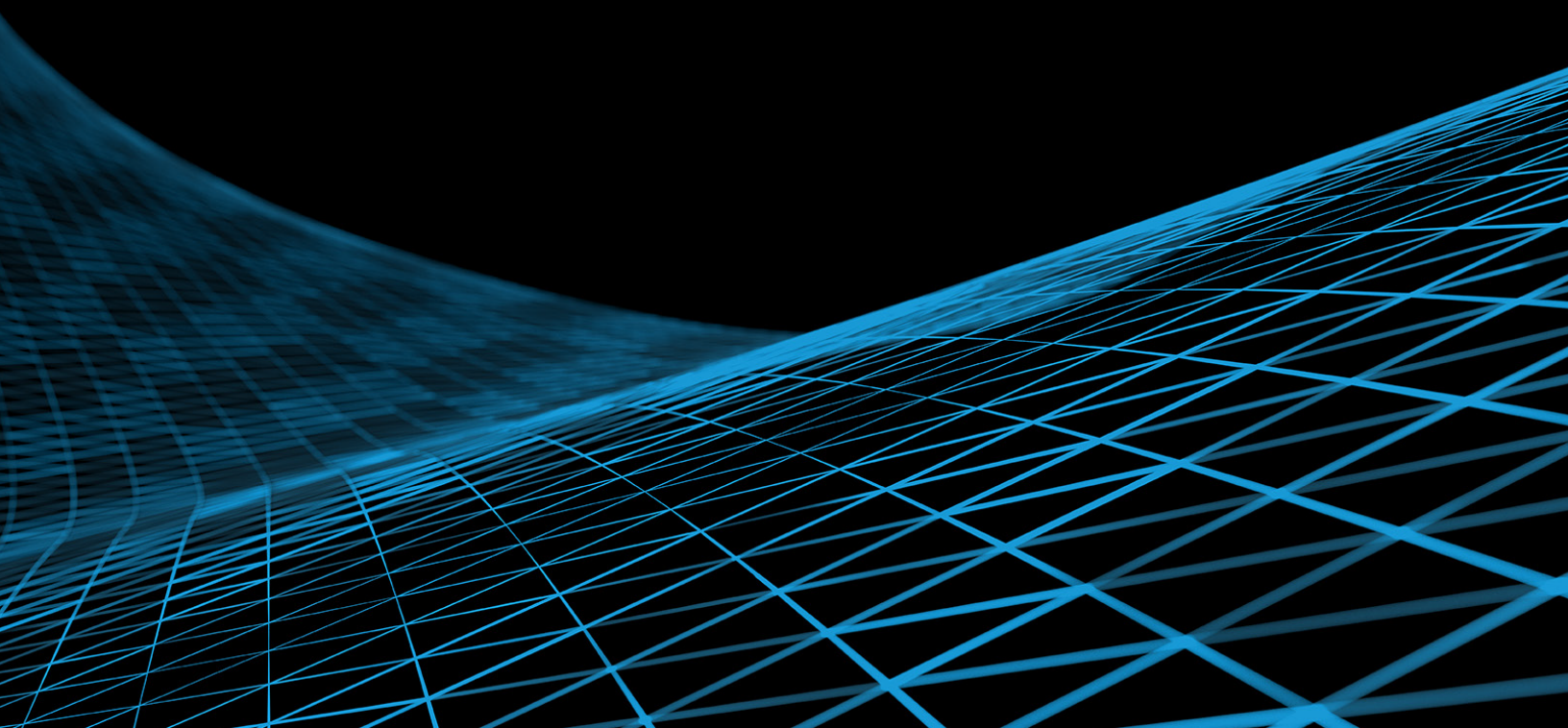


Table des matières



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Introduction

Pourquoi ce livre numérique est important

Que vous soyez un décideur informatique, un dirigeant, une partie prenante du domaine d'activité ou un membre de conseil d'administration, vous passez probablement de plus en plus de temps à discuter de l'IA, à vous informer à son sujet et à réfléchir à son impact sur votre entreprise, ainsi, sans doute, qu'à ses conséquences sur votre vie personnelle et professionnelle.

L'IA s'est développée sur le marché beaucoup plus rapidement que la plupart des technologies en vogue avant elle. Elle est l'une des rares technologies importantes de ces dernières années pour laquelle le battage médiatique initial était justifié et des résultats tangibles et observables ont été obtenus. (Nous vous donnerons quelques chiffres illustrant cette réalité.)

Cependant, ce livre numérique ne se contente pas d'inscrire la croissance et le potentiel de l'IA dans un contexte réel. Il aborde l'IA sous un autre angle important : celui des utilisateurs à proximité et du cloud distribué. Nous avons beaucoup entendu parler de l'importance de la présence des données à proximité des utilisateurs et dans le cloud. Or, nous sommes entrés dans l'ère de l'inférence de l'IA décentralisée et de l'IA dans le cloud distribué

Ce qui nous ramène au titre : Pourquoi ce livre numérique est important. Voici quatre raisons pour lesquelles les décideurs, les influenceurs et les gardiens devraient le lire :

1. **L'Edge Computing et le Cloud Computing sont les nouvelles frontières de l'innovation et de la création de valeur dans le domaine de l'IA.** En créant et en déployant des capacités d'IA en dehors du centre de données, les entreprises peuvent récolter les fruits de nouvelles connaissances, de modèles d'affaires et de solutions qui suivent le même chemin que celui emprunté par le marché de l'informatique au sens large au cours de la dernière décennie.
2. **La réglementation et la gouvernance de l'IA sont un casse-tête géant.** Si l'utilisation et les pratiques « d'IA responsables » vous semblaient complexes lorsque l'IA était déployée principalement dans des sandboxes sur site, qu'en sera-t-il des problèmes de conformité, de protection des données et de sécurité que pose l'IA à proximité des utilisateurs et dans le cloud distribué ! Votre entreprise doit exercer un contrôle strict dans ce domaine, ou elle risque de rencontrer de nombreuses difficultés.
3. **L'impact de l'IA sur l'évolution des modèles et des pratiques de travail sera encore plus important que prévu.** S'il ne fait guère de doute qu'un très grand nombre de fonctions informatiques actuellement assurées par du personnel humain vont être reprises par l'IA, le fait de sortir l'IA du centre de données va créer des opportunités passionnantes, voire inspirantes, pour les informaticiens comme pour les salariés des entreprises.
4. **Vous n'avez pas de temps à perdre.** Les initiatives d'IA sont placées tout en haut de la liste des priorités stratégiques de toutes les organisations. Si votre entreprise n'a pas encore totalement exploité l'IA sur site pour tester de nouveaux cas d'utilisation et mieux comprendre ses capacités, vous avez sans aucun doute un train de retard. Mais le mouvement vers l'IA à proximité des utilisateurs et dans le cloud distribué représente une nouvelle chance.

Une raison de plus pour laquelle ce livre numérique vous sera utile : la dernière partie offre des conseils exploitables qui permettront à votre entreprise de tirer pleinement parti de la transition vers l'IA dans le cloud distribué.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

Chapitre 1

La révolution de l'IA et comment elle a déjà changé la dynamique du marché

Si l'IA existe depuis des décennies, est-il juste de qualifier le moment actuel de « révolution de l'IA » ? Ou bien l'IA évolue-t-elle simplement, bien qu'à un rythme accéléré, à partir de ses formes et fonctions antérieures ? Ce ne sont pas des questions rhétoriques : notre secteur d'activité a tendance à présenter les révolutions techniques comme faisant irruption du jour au lendemain, sans que personne ne s'en aperçoive, à l'exception des experts techniques qui les ont déjà expérimentées en laboratoire.

Toutefois, nous pouvons qualifier la période actuelle de révolution de l'IA pour diverses raisons. Tout d'abord, il est indéniable que la croissance du marché de l'IA est époustouflante, tout comme l'importance que les dirigeants des organisations lui accordent désormais. Sa croissance et son impact actuel sur le marché sont nettement supérieurs à ceux des changements technologiques longtemps considérés comme des jalons dans le développement de l'industrie informatique (ordinateurs personnels, réseaux locaux, cloud computing et IoT, notamment).

Les données suivantes illustrent l'impact, l'importance et le potentiel de la révolution de l'IA :

- Les dépenses mondiales en matériel, logiciels et services liés à l'IA ont dépassé les 600 milliards de dollars à la fin de 2024, et monteront en flèche pour atteindre plus de 2700 milliards de dollars d'ici 2032, soit un taux de croissance annuel moyen de plus de 20 %.[1]
- 97 % des chefs d'entreprise ont déclaré que leurs entreprises bénéficient d'un rendement positif de leurs investissements dans l'IA.[2]
- Le pourcentage de PDG qui placent l'IA parmi leurs deux principales priorités technologiques est passé de 4 % à 24 % en l'espace d'un an seulement.[3]
- Comme le montre la figure ci-dessous, au cours des deux dernières années, près de la moitié des entreprises américaines ont déclaré que l'IA prenait nettement plus d'importance pour l'avenir de leur organisation ; seule la cybersécurité la devance en tant qu'initiative d'entreprise.[4]

Figure 1. La sécurité devant l'IA, le cloud et la gestion des données dans le classement des initiatives technologiques cruciales

Parmi les initiatives technologiques générales suivantes, lesquelles sont devenues beaucoup plus importantes pour l'avenir de votre entreprise au cours des deux dernières années ? (Pourcentage de répondants, N = 1, 351, sauf réponses multiples)

1 Source : Fortune Business Insights, "Artificial Intelligence market size ... 2024-2032," décembre 2024.
2 Source : Lizzie McWilliams, "Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks," EY.com, 10 décembre 2024.
3 Source : Gartner's 2024 CEO survey reveals AI as top strategic priority, LinkedIn, mai 2024.
4 Source : Enterprise Strategy Group Complete Survey Results, *2025 Technology Spending Intentions Survey*, décembre 2024.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

En réalité, la hausse de l'adoption et des investissements dans l'IA a même dépassé celle du Cloud Computing, l'un des segments les plus vigoureux et à la croissance la plus rapide du paysage informatique. La raison pour laquelle la croissance de l'IA éclipse celle du cloud est assez simple : alors que le Cloud Computing représente un moyen puissant et hautement exploitable de développer et de fournir des services informatiques, l'IA modifie fondamentalement notre façon de travailler, de vivre et d'interagir dans presque tous les aspects de notre vie.

En outre, l'adoption rapide de l'IA a révélé les limites des architectures de Cloud Computing traditionnelles. Par exemple, les modèles de cloud centralisés ont du mal à répondre aux exigences de performances intenses de l'IA en temps réel, qui nécessite énormément de données. En substance, l'IA a considérablement modifié les attentes qui pèsent sur les architectures cloud en matière de performances, de coûts, de compétences humaines, de sécurité et de gouvernance des données.

Qu'est-ce qui stimule la croissance stratosphérique et l'impact sur le marché de l'IA ? Un facteur majeur est l'utilisation beaucoup plus précise et sophistiquée des modèles d'IA par les organisations au cours de l'année écoulée. En raison de l'augmentation rapide de l'utilité et de la facilité d'utilisation des modèles d'IA, ainsi que du retour sur investissement impressionnant dont ont bénéficié de nombreuses entreprises, des cas d'utilisation significatifs ont émergé. Ces derniers renforcent la réflexion initiale sur la façon d'utiliser l'IA et ouvrent de nouveaux courants de pensée sur les possibilités de tirer profit des modèles d'IA.

Les points suivants, par exemple, doivent être pris en considération :

- **Les prévisions commerciales** ont toujours joué un rôle essentiel dans la prospérité des entreprises, et l'IA a rendu cette fonction plus précise, plus exacte, plus opportune et plus facilement exploitable. La prédiction des tendances avec des outils tels que les modèles d'IA prédictive est encore plus avancée qu'il y a dix ans, quand les entreprises adoptaient l'analyse du Big Data, car les modèles d'IA d'aujourd'hui sont boostés par les données des organisations elles-mêmes. Cela permet aux dirigeants d'entreprise, et pas seulement aux experts en science des données, d'analyser d'énormes volumes de données plus rapidement et plus précisément qu'avec les modèles traditionnels.
- **La synthèse de documents** était autrefois réservée aux travailleurs en col blanc, mais l'intelligence artificielle générative (GenAI) a changé les règles du jeu. Les longs documents (y compris les documents très techniques et ou spécialisés) sont examinés, analysés et résumés automatiquement en un temps record. Cela permet non seulement d'obtenir de meilleures informations, mais aussi d'utiliser les professionnels des services informatiques et commerciaux de manière plus innovante et stratégique, améliorant ainsi la productivité et l'expérience des collaborateurs.
- **La modélisation de l'attrition** est désormais un élément essentiel de la gestion de la relation client. Elle aide les entreprises à prendre des décisions plus pertinentes et contextuelles, pour prévoir comment, quand et pourquoi les clients pourraient se tourner vers d'autres fournisseurs, ou au contraire s'engager à augmenter leurs achats.
- **Les chatbots basés sur l'IA** transforment le service client, la gestion des services informatiques et les ressources humaines en permettant un libre-service intelligent et en temps réel pour de nombreuses demandes courantes. Ces chatbots sont même utilisés pour des cas d'utilisation d'ingénierie et de science informatique hautement techniques, tels que la génération de code et la modélisation de nouvelles fonctionnalités.

Ces cas d'utilisation ont contribué à générer de nombreuses autres applications, workflows et cas d'utilisation de l'IA, tels que la détection des menaces de cybersécurité, l'analyse des données/la veille stratégique, l'analyse concurrentielle, le prototypage rapide, la gestion de la conformité et bien plus encore.

Bien sûr, toute nouvelle technologie (même une technologie aussi dynamique et porteuse de promesses que l'IA) pose des défis en matière de développement, de déploiement et de gestion. Ces défis peuvent être techniques, par exemple la qualité insuffisante des données, la nécessité de créer une infrastructure adaptée ou la garantie des hautes performances nécessaires pour piloter de grands modèles de langage (LLM). Ils peuvent également être liés à l'activité ou aux risques, tels que la gouvernance de données, la garantie d'une utilisation responsable de l'IA ou la résolution de l'écart de compétences important et croissant en matière d'IA.

Les premières indications montrent que les organisations comprennent et relèvent ces défis avec les bonnes ressources et le soutien des cadres dirigeants et du conseil d'administration. Dans le chapitre suivant, nous expliquerons l'une des principales évolutions qui font de l'IA plus qu'un projet scientifique : le passage de l'entraînement de modèles à l'inférence.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh
```

Chapitre 2

Le changement de paradigme de l'IA : de l'entraînement à l'inférence

L'entraînement de modèles d'IA a été un facteur essentiel du développement du marché et se trouvait au cœur des initiatives des entreprises en matière d'IA. Les entreprise ont vite compris que le manque de qualité des données (ou des données inappropriées ou en quantité insuffisante) affectait le développement des modèles d'IA. Une étude menée par l'Enterprise Strategy Group d'Informa TechTarget a révélé que 37 % d'entre elles citent les problèmes de qualité des données comme leur principal défi dans la mise en œuvre de la GenAI, ce qui en fait le deuxième défi le plus mentionné après le manque de compétences en matière d'IA.[5]

Les entreprises ayant pris conscience que le meilleur moyen de résoudre les problèmes de qualité des données et de créer des LLM plus précis, plus contextuels et plus réactifs consistaient à puiser dans leurs propres données, elles peuvent désormais construire plus facilement le bon modèle d'IA. Cette approche donne souvent de meilleurs résultats que l'achat de modèles d'IA prêts à l'emploi entraînés à partir de données tierces, qui peuvent créer des distorsions et des inexactitudes dans les données et offrent peu de capacités de personnalisation. En pratique, la plupart des entreprises ont concentré l'entraînement de leur LLM dans des environnements cloud centralisés, une approche efficace pour la première vague de développement LLM.

L'entraînement des modèles d'IA a évolué et s'est stabilisé en termes de qualité et d'intégrité. Déjà, les entreprises allouent moins de ressources à leurs modèles d'entraînement internes, en particulier dans la mesure où le réglage et l'optimisation de l'IA réduisent le besoin d'une infrastructure axée sur le processeur graphique et nécessitant un traitement intensif.

Mike Leone, directeur des pratiques d'IA, d'analyse et de business intelligence chez Enterprise Strategy Group, offre un point de vue indépendant sur cette transition de l'apprentissage vers l'inférence :

« À mesure que les modèles pré-entraînés et les applications d'IA en temps réel gagnent en maturité et en importance, l'inférence joue désormais un rôle central dans la réalisation efficace et rentable de la valeur promise par l'IA », affirme-t-il. « Ce changement de priorité permet des déploiements à grande échelle dans divers secteurs, une tendance favorisée par les progrès continus réalisés à la fois en termes de matériel et de logiciel. Les fournisseurs se concentrent sur le dimensionnement et l'optimisation de l'infrastructure d'IA, l'amélioration de l'efficacité des modèles et la gestion des coûts opérationnels continus. Bien sûr, alors que l'inférence démocratise l'accès à l'IA, elle introduit des défis tels que l'évolutivité, les préoccupations en matière de protection de la vie privée et la conformité réglementaire. »

La transition de l'entraînement de modèles d'IA vers des applications d'IA basées sur l'inférence et des cas d'utilisation à fort impact est en bonne voie. Armand Ruiz, vice-président d'IBM pour les plateformes d'IA, l'exprime de manière concise et forte : « À mon avis, l'inférence dominera les charges de travail de l'IA. ... Une fois qu'un modèle a été correctement entraîné, il n'a idéalement plus besoin d'entraînement supplémentaire. L'inférence, en revanche, est continue ».[6]

Lorsque des transitions de marché ont lieu, elles sont inévitablement marquées par des variations disproportionnées du mode et de la destination des investissements en capital. Par exemple, une estimation a souligné qu'environ 15 % des dépenses en puces de silicium pour l'IA sont consacrées à l'entraînement de l'IA, 45 % allant à l'inférence de l'IA basée sur les centres de données et les 40 % restants à l'inférence en bordure de l'Internet.[7]

5 Source : Enterprise Strategy Group Complete Survey Results, "The State of the Generative AI Market: Widespread Transformation Continues," Septembre 2024.
6 Source : LinkedIn, Armand Ruiz, décembre 2024.
7 Source : Jonathan Goldberg, "The AI chip market landscape: Choose your battles carefully," TechSpot.com, 30 mai 2023.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

D'autres recherches ont mis en évidence une tendance similaire en faveur de l'inférence. Une étude de Google révèle notamment que 60 % de sa consommation d'énergie centrée sur l'apprentissage automatique est utilisée pour l'inférence, contre seulement 40 % pour l'entraînement.[8]

Il n'est pas surprenant que ce passage de l'entraînement de modèles d'IA à l'inférence de l'IA ait des implications significatives au niveau de l'infrastructure IA et des architectures cloud. Cette transformation est particulièrement importante pour les entreprises qui recherchent l'environnement et l'architecture cloud appropriés à utiliser pour le développement, le déploiement et la gestion des applications d'IA.

Les modèles de cloud centralisés traditionnels fonctionnent toujours bien pour de nombreuses charges de travail professionnelles, mais l'IA met l'architecture cloud sous pression pour répondre aux exigences de performances de l'IA à grande échelle. Ces défis se manifesteront probablement sous la forme de problèmes de latence frustrants, avec un délai notable entre les requêtes et les réponses.

L'infrastructure informatique est également un élément clé dont les entreprises doivent tenir compte lorsqu'elles passent de l'entraînement de modèles d'IA à l'inférence de l'IA. Les processeurs graphiques centrés sur l'IA ont été plébiscités pour leur capacité à fournir la puissance nécessaire à l'entraînement des LLM, car ils offrent des performances de calcul et une bande passante mémoire élevées. En comparaison, l'inférence nécessite une approche différente en matière d'infrastructure, qui doit être axée sur une faible latence et une utilisation efficace des ressources.

Dans le chapitre suivant, nous examinerons de plus près comment les entreprises peuvent contribuer à répondre aux exigences de performance de l'inférence de l'IA en positionnant les ressources d'inférence au plus près de l'utilisateur et dans une architecture cloud distribuée.

8 Source : "*The AI boom could use a shocking amount of electricity.*" Scientific American, 13 octobre 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Chapitre 3

L'essor de l'inférence de l'IA décentralisée et du cloud distribué

Comme l'explique le chapitre précédent, l'architecture cloud centralisée traditionnelle impose des limitations et des défis problématiques aux entreprises qui cherchent à optimiser leur utilisation de l'IA dans la création d'applications révolutionnaires et transformatrices. Des questions telles que la latence, les limites de la bande passante du réseau, la garantie des performances à grande échelle et la complexité de la gestion, doivent toutes être prises en compte et planifiées, à mesure que les exigences de l'IA passent de l'entraînement de modèles à l'inférence.

Les applications d'inférence de l'IA actuelles et de nouvelle génération nécessitent de pouvoir développer, déployer et exploiter l'inférence là où résident les données. De plus en plus, cela se traduit par une proximité avec les utilisateurs ou par une architecture cloud distribuée agile, évolutive et flexible.

Avant d'approfondir, rappelons quelques définitions utiles.

- *L'inférence en IA décentralisée* est un paradigme permettant d'élaborer des flux de travail d'IA qui couvrent les centres de données centralisés (le cloud) et les terminaux en dehors du cloud qui sont plus proches des humains et des objets physiques (la bordure de l'Internet). Cela contraste avec la pratique la plus courante consistant à développer et exécuter les applications d'IA entièrement dans le cloud, aussi appelée l'*IA dans le cloud*. Cela diffère également des anciennes approches de développement de l'IA qui consistaient à créer des algorithmes d'IA sur des ordinateurs de bureau, puis à les déployer sur des ordinateurs de bureau ou du matériel spécial pour des tâches telles que la lecture de numéros de chèques.

La diffusion de services cloud peut aider les entreprises à atteindre leurs objectifs en matière de performances applicatives et de temps de réponse, d'Edge Computing, de réglementations ou d'autres exigences...

- Akamai, l'un des principaux fournisseurs de plateformes pour le cloud, la sécurité et la diffusion de contenu, définit le *cloud distribué* comme « une forme de Cloud Computing dans laquelle une entreprise utilise une infrastructure de cloud public répartie sur plusieurs sites géographiques, tandis que les opérations, la gouvernance et les mises à jour sont gérées de manière centralisée par un seul fournisseur de services de cloud public. La distribution des services cloud peut aider les entreprises à atteindre leurs objectifs en matière de performances des applications et de temps de réponse, d'Edge Computing, de réglementations ou d'autres exigences en fournissant des services cloud au plus près des utilisateurs et des terminaux. L'utilisation du Cloud Computing distribué est motivée par l'essor des réseaux Internet des objets (IoT), de l'intelligence artificielle et d'autres cas d'utilisation qui nécessitent le traitement de grandes quantités de données en temps réel. »

L'inférence de l'IA décentralisée et le cloud distribué sont essentiels au développement et à l'utilisation de l'inférence de l'IA, car ils représentent des moyens évolués, efficaces, sécurisés et fiables de surmonter les défis mentionnés précédemment : hautes performances, performances à grande échelle, utilisation efficace des ressources de calcul, de stockage et de réseau, ainsi que résolution des problèmes de latence associés aux grands ensembles de données, souvent constitués d'énormes quantités de données non structurées.

Comment cela fonctionne-t-il dans la vie réelle ? Prenons pour exemple une entreprise qui souhaitait faire de la personnalisation le point fort de son expérience client et devait s'assurer qu'elle capturait des données provenant de l'ensemble des terminaux et des sources de données. Elle a décidé d'utiliser un chatbot IA basé sur un LLM pour distribuer des données localisées à des emplacements proches des utilisateurs. Cela a permis de réduire la latence et d'améliorer les temps de réponse, et ainsi d'améliorer considérablement les performances et de fournir un support client plus rapide et plus personnalisé.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

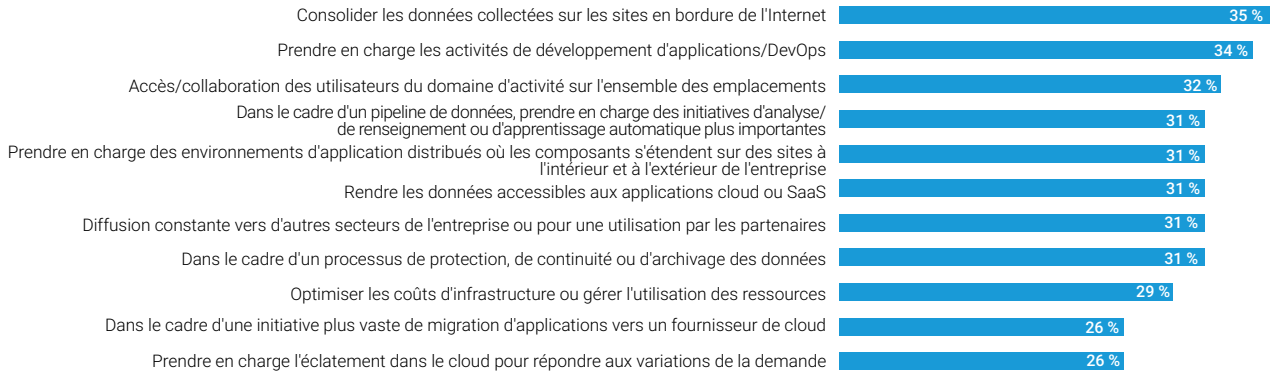

De plus en plus d'entreprises adoptent cette approche pour tirer parti des opportunités croissantes offertes par les applications d'inférence de l'IA. Pour résoudre les problèmes de performance liés aux limitations de l'infrastructure de calcul et aux problèmes de latence, les charges de travail d'IA se rapprochent des sources de données qu'un plus grand nombre d'utilisateurs finaux utilisent via l'inférence de l'IA décentralisée et les architectures cloud distribuées.

Par exemple, ce modèle est largement utilisé dans les secteurs marqués par des installations très distribuées et géographiquement dispersées, notamment le commerce de détail, les télécommunications et les médias. De plus en plus, cette approche devient le modèle architectural standard pour l'inférence de l'A en raison de sa capacité à surmonter les défis de latence et de performances à grande échelle associés aux exigences complexes de l'inférence de l'IA.

L'un des principaux facteurs de passage à un modèle de cloud distribué pour l'inférence de l'IA est la nécessité croissante pour les entreprises d'utiliser la bordure de l'Internet pour déplacer des données entre les centres de données et les fournisseurs de services cloud (CSP). Enterprise Strategy Group a constaté que 35 % des entreprises qui déplacent régulièrement des données entre les centres de données et l'infrastructure des CSP le font pour consolider les données collectées en bordure de l'Internet ; il s'agit de la principale motivation de ce mouvement de données.[9]

Figure 2. La bordure de l'Internet, principal moteur de la circulation des données entre les centres de données et les services de cloud public

Vous avez indiqué que votre entreprise déplace régulièrement des données entre son ou ses centres de données et les services de cloud public. À quelles fins votre organisation transfère-t-elle des données entre ses centres de données et les services de cloud public ? (Pourcentage de personnes interrogées, N = 170, mult)



9 Source : Enterprise Strategy Group Research Report, Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud, novembre 2023.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

Chapitre 4

Raisons pour lesquelles l'inférence de l'IA décentralisée représente l'avenir

Construire et déployer l'inférence de l'IA avec des données provenant de sites en bordure de l'Internet est essentiel pour aider les entreprises à tirer le meilleur parti de l'inférence. Compte tenu de la hausse de la création et du stockage de données sur divers appareils en bordure de l'Internet (par exemple, smartphones, terminaux portables spécialisés, dispositifs de signalisation digitale et IoT), les entreprises doivent adopter un esprit orienté sur la bordure de l'Internet pour le traitement des données en temps réel indispensable à l'inférence de l'IA.

Le passage à l'inférence de l'IA décentralisée pour s'aligner sur les exigences de l'inférence de l'IA a été rapide et profond. Les études indiquent que les dépenses consacrées aux systèmes d'inférence proches des utilisateurs seront multipliées par dix, passant de 27 milliards de dollars en 2024 à près de 270 milliards de dollars en 2032.[10]

L'inférence de l'IA décentralisée est marquée par un certain nombre de considérations importantes pour aider les entreprises à développer et à déployer des applications d'inférence de l'IA, telles que le traitement IA natif des terminaux, les cadres de sécurité intégrés optimisés pour les terminaux en bordure de l'Internet, la faible consommation d'énergie et la faible latence. Ces fonctionnalités se traduisent par des performances plus rapides à grande échelle, des informations plus précises et plus approfondies sur les données, ainsi que la réduction des investissements en bande passante. Cela permet ensuite aux membres d'une entreprise dispersés géographiquement d'utiliser l'inférence de l'IA pour prendre des décisions plus intelligentes en temps réel, beaucoup plus près de là où les données sont créées, stockées, traitées et consultées.

Bloomfield est un exemple d'entreprise avant-gardiste ayant utilisé l'inférence de l'IA pour améliorer ses opérations et faciliter la gestion des données. Elle devait gérer de très grands volumes de données provenant de ses caméras intelligentes distribuées de manière efficace, économique et sécurisée pour les agriculteurs, qui disposent souvent d'un accès réseau limité. En utilisant un modèle architectural d'inférence de l'IA décentralisée, Bloomfield a tiré profit de l'interférence de l'IA pour mieux déterminer la meilleure façon d'organiser les données, de les stocker et de les présenter aux agriculteurs. Cela a permis de simplifier la gestion de la santé des cultures, d'améliorer les rendements de production et de prendre des décisions plus rapides et plus précises loin d'un centre de données sur site.

La capacité de l'inférence de l'IA décentralisée à prendre en charge des applications d'inférence de l'IA dans une variété de sites distribués le rend naturellement propice à des cas d'utilisation tels que :

- **Médecine de chevet et applications de santé à distance,** où les décisions sont prises par des cliniciens en déplacement à l'aide de smartphones, de terminaux IoT et d'autres terminaux portables légers mais puissants. L'inférence de l'IA décentralisée est également idéale pour prendre des décisions en temps réel à l'aide d'appareils intelligents dans le secteur de la santé, tels que les pompes à perfusion, les stimulateurs cardiaques, les moniteurs cardiaques et l'analyse des médicaments.
- **Villes intelligentes,** où les systèmes en bordure de l'Internet font tout, depuis le contrôle des flux de circulation et la surveillance du comportement des systèmes de services d'utilité publique jusqu'à la rationalisation de l'inscription des électeurs et l'amélioration de la sécurité publique.
- **Véhicules sans conducteur,** qui utilisent des caméras intelligentes et des données de capteurs embarqués pour évaluer les options d'itinéraire, surveiller et améliorer la consommation de carburant, documenter les changements de GPS et communiquer avec les forces de l'ordre et les organismes de sécurité publique.
- **Commerce de détail,** qui prend en charge un large éventail d'applications d'inférence de l'IA, de la vidéosurveillance IP et la gestion des stocks à la prévention des vols et aux schémas de circulation des clients dans les magasins.

Alors que les entreprises cherchent des moyens de réduire la latence des réponses aux requêtes en temps réel, l'inférence de l'IA décentralisée deviendra une exigence cruciale pour une source de prise de décision plus efficace, mieux informée et plus précise. Les applications d'IA n'auront plus besoin d'envoyer des données vers des centres de données éloignés ou de demander des données à partir de ces centres : elles traiteront et partageront des données sur un cloud distribué dans le monde entier.

Le responsable de l'ingénierie d'une grande entreprise de technologie publicitaire déclare : « Nous avons passé beaucoup de temps à optimiser notre modèle d'IA. Aujourd'hui, nous réalisons que l'inférence ne doit pas être une réflexion après coup ».

Les entreprises ne peuvent plus s'offrir le luxe de décider si elles doivent adopter l'inférence de l'IA comme un élément majeur de leur stratégie d'entreprise. Elles doivent prendre conscience de l'urgence et utiliser l'inférence de l'IA décentralisée pour réaliser l'inférence, car nombre de leurs concurrents ont déjà fait de grands pas vers l'utilisation de l'inférence de l'IA à proximité des utilisateurs.

10 Source : Fortune Business Insights, « Edge AI Market Size … 2024-2032, » décembre 2024.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Chapitre 5

Aperçu stratégique de l'IA, de la bordure de l'Internet et du cloud distribué : où allons-nous et qu'est-ce que cela signifie

Évaluer et sélectionner un partenaire technologique expérimenté et collaborer avec lui est un élément fondamental, mais délicat, de la réussite d'une initiative d'inférence de l'IA d'une entreprise.

Fondamental, car il est indéniable qu'il existe un déficit substantiel de compétences en matière d'IA, que ce soit au niveau de l'infrastructure, de l'architecture cloud, des cas d'utilisation de l'inférence, de l'optimisation des applications ou de la perfection du déploiement.

Mais délicat, en raison du manque de candidats appropriés. Comme nous l'avons vu plus haut, un projet d'inférence de l'IA réussi doit combiner le savoir-faire technique, l'expertise sectorielle, la compréhension de la sécurité, de la gouvernance et de la gestion des risques, ainsi qu'une attention rigoureuse aux détails lors de la mise en œuvre et de la gestion continue.

Pour développer et déployer l'inférence de l'IA dans l'ensemble de l'entreprise physique et virtuelle, du cœur au cloud en passant par la bordure de l'Internet, les entreprises doivent se rapprocher d'un fournisseur qui l'a déjà fait par le passé. Il est essentiel de disposer d'un partenaire qui comprend à la fois les éléments techniques et commerciaux d'un projet d'IA réussi.

Akamai, un fournisseur majeur de solutions sophistiquées pour le Cloud Computing, la sécurité et la diffusion de contenu, a une solide expérience dans la construction, la mise en œuvre et la gestion d'initiatives d'intelligence artificielle, allant de l'entraînement de modèles à l'inférence. Son expérience pratique des charges de travail complexes et gourmandes en ressources de calcul l'a préparé à travailler avec un large éventail de clients, à travers différents cas d'utilisation, secteurs d'activité et zones géographiques.

Pour aider les entreprises à répondre facilement à leurs exigences en matière d'inférence de l'IA, Akamai a déployé avec succès, de manière cohérente, fiable et sécurisée, une architecture cloud distribuée permettant aux entreprises de traiter les charges de travail d'IA au plus près de l'utilisateur final. Cela s'est avéré inestimable pour réduire la latence et assurer des performances élevées à grande échelle, deux éléments essentiels pour les charges de travail d'inférence de l'IA à forte intensité de calcul.

L'un des principaux avantages de l'architecture cloud distribuée d'Akamai est la capacité pour les développeurs de se concentrer exclusivement sur le développement de plateformes neutres dans une infrastructure basée sur le cloud et fondée sur des normes de cloud ouvertes.

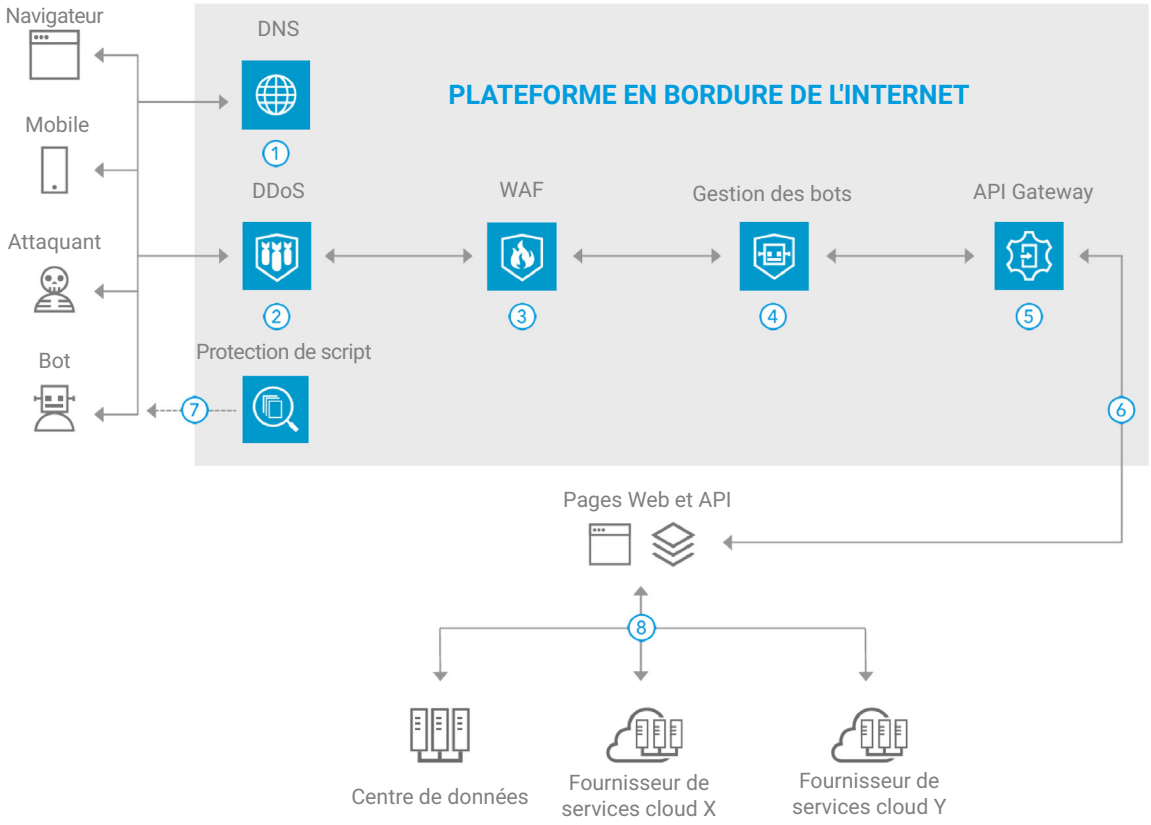
Pour atteindre des performances élevées à grande échelle, Akamai utilise plusieurs techniques de mise en cache différentes, intégrées à des logiciels open source, afin de garantir un débit suffisant et de limiter la latence. Par exemple, de nombreux clients d'Akamai utilisant son architecture cloud distribuée ont déclaré avoir obtenu une réduction de 50 % des temps de réponse pour les chatbots en contact avec les clients, un cas d'utilisation très important et très populaire de l'inférence. Ce type de réponse en temps quasi réel permet aux entreprises de nombreux secteurs, notamment le commerce de détail, la santé, les services financiers et la haute technologie, de fournir un service client plus rapide et mieux adapté à leurs besoins spécifiques.

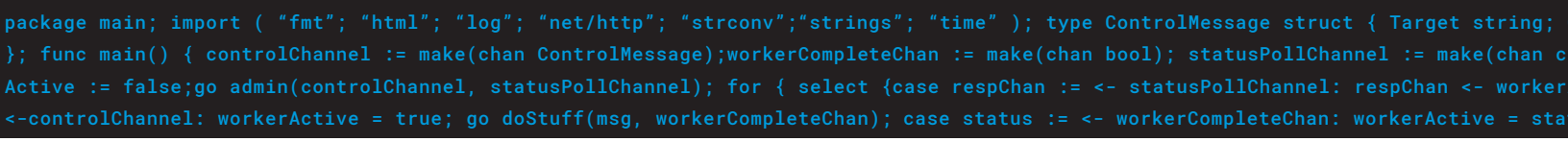
```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Akamai propose également une architecture de référence d'IA (voir la figure 3 ci-dessous) qui permet aux entreprises de démarrer rapidement et d'obtenir une valeur économique et opérationnelle élevée. Cette architecture de référence est complète dans la mesure où, grâce à un ensemble familier de navigateurs et de terminaux définis par l'utilisateur, les entreprises peuvent déployer une solution d'IA d'entreprise depuis le centre de données central jusqu'au cloud en passant par la bordure de l'Internet, et inversement. Cela profite à tous les acteurs de l'entreprise : développeurs, spécialistes informatiques, responsables d'infrastructure, architectes de cloud et parties prenantes du domaine d'activité.

Figure 3.

Avec les progrès rapides de la technologie et la complexité et la valeur croissantes des charges de travail d'IA, les entreprises doivent agir rapidement et de manière décisive pour surpasser la concurrence. Il est indispensable d'utiliser l'inférence de l'IA à son potentiel maximal en adoptant l'inférence de l'IA décentralisée et une architecture cloud distribuée, en particulier planifiées, développées, déployées et gérées par un leader établi comme Akamai.





Conclusion

Les prochaines étapes pour votre entreprise

On dit que l'IA a changé les règles du jeu en matière d'excellence informatique et de transformation de l'entreprise. Cependant, la clé pour optimiser la valeur des investissements dans l'IA d'une entreprise consiste à passer rapidement de l'entraînement de modèles d'IA à l'inférence de l'IA.

Pour tirer le meilleur parti de tout ce que l'inférence de l'IA a à offrir, les entreprises doivent développer et déployer leurs applications d'inférence et leurs charges de travail au plus près de l'endroit où le travail est effectué ; le personnel clé pourra ainsi réaliser des choses remarquables avec les systèmes. Cela signifie passer d'un modèle de Cloud Computing centralisé à un modèle de cloud distribué fortement façonné par l'IA en bordure de l'Internet.

Comme indiqué précédemment, les entreprises doivent prendre en compte de nombreux problèmes et relever un certain nombre de défis. Celles qui n'ont pas encore placé les initiatives d'IA parmi leurs priorités stratégiques peuvent se sentir dépassées, mais il existe des étapes importantes qu'elles peuvent et doivent mettre en œuvre pour aller de l'avant et ne pas être distancées.

« Les entreprises qui rattrapent leur retard en matière d'IA ne doivent pas laisser leur peur de rater le coche de cette technologie les piéger », souligne Mike Leone, de l'Enterprise Strategy Group. « L'accent doit être mis sur la création des bonnes bases en prenant des mesures telles que l'identification des problèmes métiers spécifiques que l'IA peut résoudre, l'évaluation de leur infrastructure de données existante et de leurs compétences internes, et l'identification des partenaires clés aptes à combler les lacunes. Elles doivent faire preuve de stratégie en identifiant un ou deux projets pilotes à fort impact et des cas d'utilisation qui peuvent démontrer une valeur évidente en utilisant des données de qualité existantes. »

Pour qu'une initiative d'inférence de l'IA soit couronnée de succès, il est essentiel de savoir ce qui devra être mesuré, et comment.

« En parallèle, elles doivent également développer les talents internes et mettre en place des cadres de gouvernance pour garantir une utilisation responsable. L'essentiel est d'éviter la mise en œuvre précipitée au profit d'étapes stratégiques qui créent des capacités fondamentales et, idéalement, durables. »

À cette fin, voici quatre règles que vous et vos collègues devez comprendre et vous engager à respecter au plus vite :

1. **Sélectionnez les cas d'utilisation avec soin.** Il s'agit d'une première étape essentielle, car vous devrez prendre confiance dans votre capacité à déployer correctement l'inférence de l'IA. Il est également important de renforcer la confiance entre les parties prenantes de votre entreprise et les approbateurs financiers grâce à quelques bénéfices précoces.
2. **Veillez à bien choisir l'infrastructure.** L'inférence de l'IA ayant une forte intensité de calcul, la première option venant à l'esprit est d'adopter le processeur graphique leader sur le marché pour le cœur de votre système. Mais il y a beaucoup d'autres excellents choix de produits performants pour les processeurs graphiques et autres infrastructures d'IA. Prenez le temps de déterminer l'infrastructure la plus appropriée (vous pouvez aussi faire appel à l'expérience et à l'expertise d'un partenaire expérimenté en la matière).
3. **Obtenez l'adhésion de la direction.** Concentrez-vous sur les deux préoccupations communes aux dirigeants et aux membres du conseil d'administration : quelle est l'opportunité pour nous, et comment trouver un équilibre avec le risque potentiel ? Assurez-vous que votre projet et vos cas d'utilisation correspondent aux principaux objectifs de l'entreprise, tels que l'amélioration de l'expérience client, l'identification et l'exploitation plus rapides des opportunités du marché et la création d'un processus de bêta-test plus précis. Pendant ce temps, identifiez et recrutez des sponsors de haut niveau issus de groupes tels que le service informatique, la ligne métier et les membres du conseil d'administration.
4. **Définissez et assurez la réussite.** Pour qu'une initiative d'inférence de l'IA soit couronnée de succès, il est essentiel de savoir ce qui devra être mesuré, et comment. Sélectionnez des indicateurs (ou travaillez avec votre partenaire pour développer des indicateurs personnalisés) qui sont réalistes, pertinents et suffisamment solides pour constituer de « grandes victoires », en particulier après avoir collecté certains succès initiaux faciles avec les premiers cas d'utilisation de l'inférence de l'IA.

Ce contenu a été commandé par Akamai et produit par TechTarget Inc.