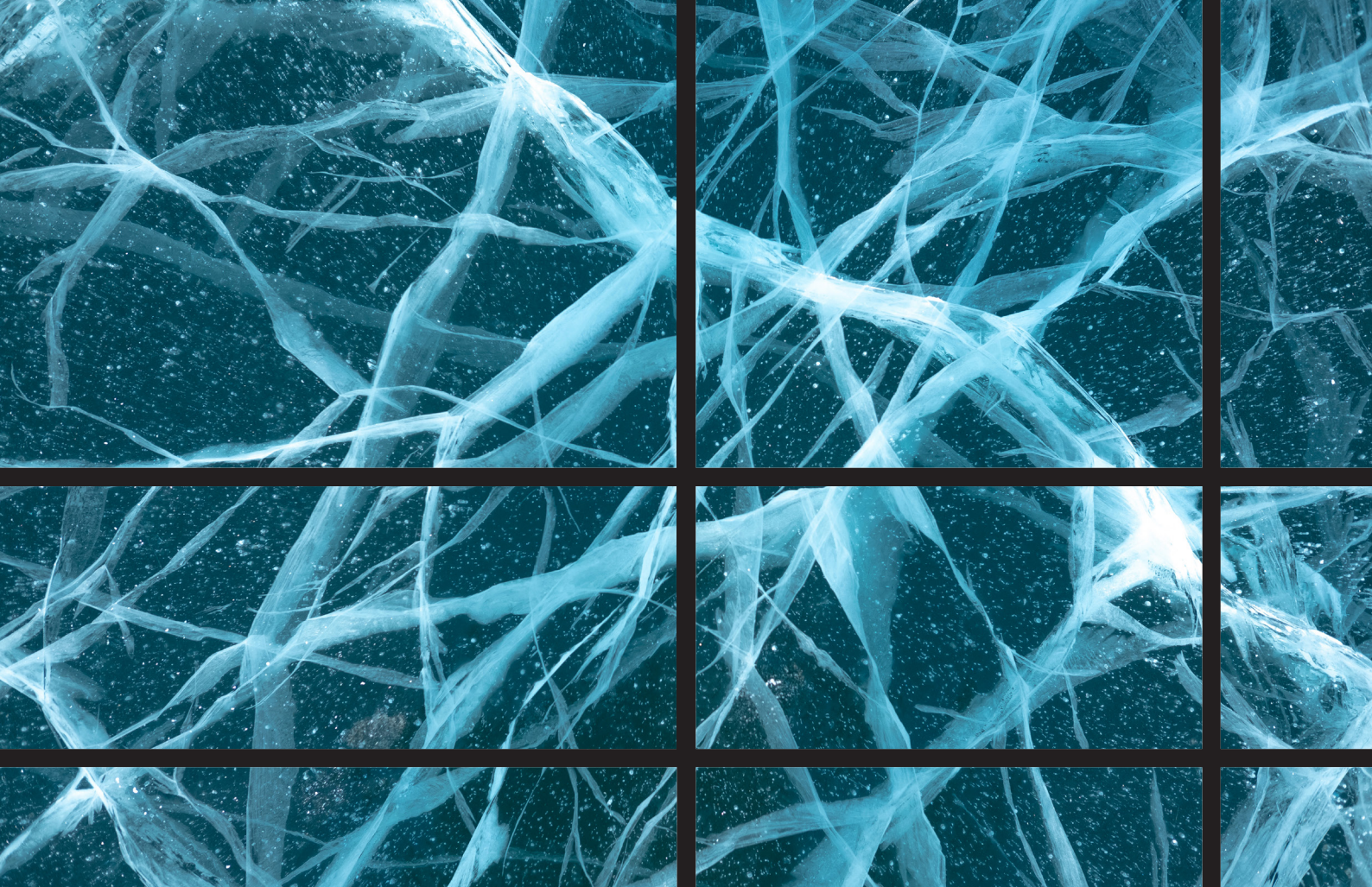




Rapport de découverte 451 Research

Mai 2025

Architectures IA à grande échelle

Les exigences réseau vues par des responsables et dirigeants

Commissioned by

S&P Global
Market Intelligence

verizon

Table des matières

Introduction	3
Méthodologie	3
Figure 1 : Répartition démographique des participants	3
Synthèse	4
Point n°1 : La planification des réseaux doit anticiper les besoins et ne pas se contenter de réagir aux cas d'usage actuels	5
Figure 2 : Domaines fonctionnels présentant les cas d'usage les plus avancés en matière d'IA/ML	6
Figure 3 : Domaines d'application de l'IA dans l'industrie, la distribution d'énergie et la santé/les sciences de la vie	7
Point n°2 : Les entreprises leaders s'appuient sur une infrastructure diversifiée pour l'IA, et les stratégies réseau doivent refléter cette réalité	8
Figure 4 : L'orchestration multicloud émerge comme un besoin essentiel pour l'infrastructure	10
Point n°3 : La latence, la bande passante et la disponibilité déterminent les stratégies réseau des entreprises les plus en pointe sur l'IA	11
Figure 5 : Principales capacités des partenaires technologiques pour l'infrastructure réseau/de sécurité des réseaux	12
Point n°4 : La gestion des workloads IA bénéficie de la segmentation du réseau	13
Figure 6 : Efficacité des différentes approches pour les workloads d'IA	14
Implications	15
Point n°1 : La planification des réseaux doit anticiper les besoins et ne pas se contenter de réagir aux cas d'usage actuels.	15
Point n°2 : Les entreprises leaders s'appuient sur une infrastructure diversifiée pour l'IA, et les stratégies réseau doivent refléter cette réalité.	16
Point n°3 : La latence, la bande passante et la disponibilité déterminent les stratégies réseau des entreprises les plus en pointe sur l'IA.	16
Point n°4 : La gestion des workloads IA bénéficie de la segmentation du réseau.	16
À propos de l'auteur	17

Introduction

L'attrait de l'IA générative a suscité l'intérêt des managers et dirigeants d'entreprise, lesquels la considèrent comme une technologie transformatrice vouée à actionner de nouveaux leviers d'efficacité et de compétitivité. Des budgets importants ont ainsi permis d'expérimenter rapidement l'IA, avec un soutien marqué des Comex. Le principal défi consiste désormais à passer de ces phases expérimentales à un déploiement opérationnel à grande échelle. Ce rapport est le premier d'une série en trois parties qui vise à recueillir les conseils de précurseurs de premier plan qui ont su piloter la mise en production de l'IA à grande échelle, en mettant l'accent sur les bonnes pratiques mises en œuvre et les enseignements qu'ils en ont tirés.

Dans ce premier volet, nous nous concentrerons sur le réseau. Nous explorerons en particulier les stratégies réseau déployées pour mener à bien les grands projets d'IA, ainsi que les obstacles rencontrés au cours de ce processus. Le deuxième rapport examinera les implications de l'IA en matière de sécurité, et le troisième abordera les bonnes pratiques globales identifiées par les porteurs de projets IA au cours de l'étude.

Méthodologie

Les informations présentées dans ce rapport sont issues d'entretiens approfondis et de discussions de groupe avec des porteurs de projets d'IA au sein d'entreprises qui ont réussi à implémenter l'IA à grande échelle. Ces échanges se sont composés de 15 entretiens approfondis avec des responsables dotés de pouvoirs décisionnaires sur les questions d'IA et d'infrastructure IT, ainsi qu'un groupe de discussion de 30 participants. Tous les participants à l'étude étaient spécifiquement impliqués dans la gestion et l'implémentation de l'infrastructure IT de leur entreprise pour les workloads d'IA. Ces sondés étaient établis aux États-Unis, au Royaume-Uni, à Singapour, en Australie, en Suède, en Allemagne, au Danemark et au Japon. Ce rapport a été réalisé pour Verizon par S&P Global Market Intelligence 451 Research.

Figure 1 : Répartition démographique des participants

Groupe de discussion				Entretiens individuels			
Région	N°	Secteur	N°	Région	N°	Secteur	N°
Amérique du Nord	10	Santé/sciences de la vie	8	Amérique du Nord	5	Retail/vente en gros	3
Europe	10	Services financiers	8	Europe	5	Santé/sciences de la vie	4
Asie-Pacifique	10	Industrie	6	Asie-Pacifique	5	Services financiers	5
		Retail/vente en gros	3			Industrie	2
		Distribution d'énergie	1			Distribution d'énergie	1
		Autre	4				

Source : Étude à grande échelle réalisée pour Verizon par S&P Global Market Intelligence 451 Research AI.

Synthèse

Les entreprises sont clairement désireuses de recourir à l'IA, et celles qui montrent la voie investissent dans un ensemble de modèles génératifs et de modèles de reconnaissance des formes et motifs. Les dirigeants, managers et opérationnels qui ont participé à l'étude sont tout à fait conscients des possibilités offertes par la technologie, notamment en termes d'efficacité et de réduction des coûts et, dans certains cas, de croissance du chiffre d'affaires.

« Nos principaux moteurs de déploiement de l'IA sont l'amélioration de l'expérience client et de l'efficacité opérationnelle, ainsi que le renforcement de la gestion des risques. »

Responsable architecture/cybersécurité

Secteur bancaire, 20 000 à 50 000 employés, Singapour

Les cas d'usage de l'IA mentionnés par les participants à l'étude sont généralement axés sur l'optimisation des tâches fonctionnelles, telles que la prévision de la demande, la conformité et l'automatisation des processus. En outre, l'amélioration des analyses de données et l'aide à la prise de décision semblent être les ressorts constants de nombreuses initiatives dans le domaine de l'IA. Cet aspect reflète les données de l'enquête de 451 Research Voice of the Enterprise: AI & Machine Learning, Use Cases 2025, laquelle a révélé que parmi les objectifs d'entreprise évalués, l'IA générative est la plus constante en termes d'efficacité opérationnelle.

Pour de nombreuses personnes interrogées, les attentes vis-à-vis de l'IA vont au-delà des capacités internes pour s'étendre aux applications en prise directe avec l'extérieur. De nombreux participants investissent notamment dans l'IA pour améliorer les interactions de leur service client. Parmi les autres cas d'usage de l'IA mis en avant par les participants figurent la rationalisation des processus de vente, l'implémentation d'un marketing personnalisé et la stimulation de l'innovation produit.

« L'automatisation pilotée par l'IA minimise les efforts manuels, réduisant ainsi les coûts opérationnels tout en améliorant la précision et la rapidité. »

Responsable IA et cloud

Services professionnels, >50 000 employés, Royaume-Uni

Il ressort clairement de l'étude que l'infrastructure constitue un obstacle majeur à la réalisation de ces objectifs, et que la capacité de calcul (p. ex. la disponibilité des GPU) n'est pas la seule contrainte matérielle entrant en ligne de compte. À mesure que les entreprises se lancent dans l'aventure de l'IA, elles prennent conscience du besoin de renforcer leurs ressources réseau. La grande majorité (90 %) des participants au groupe de discussion prévoit des changements dans leur infrastructure réseau au cours des 12 à 24 prochains mois, tandis que 71 % prévoient des mises à niveau allant de modérées à importantes. Lorsqu'on leur demande quelles sont les plus gros défauts de conception rencontrés dans leurs initiatives d'IA, de nombreuses personnes interrogées pointent du doigt les problèmes liés à l'architecture réseau.

« [Notre erreur de conception a été] de ne pas prendre en compte les exigences en matière de latence et de bande passante. Les grands modèles nécessitent des téraoctets pour ce transfert. »

Responsable architecture/cybersécurité

Secteur bancaire, 20 000 à 50 000 employés, Singapour

« L'exécution de LLM dans [notre] infrastructure actuelle ne serait pas possible en raison d'un manque de bande passante, rien que pour prendre en charge le RoCEv2. »

Architecte senior, sécurité de l'infrastructure et des réseaux

Dispositifs médicaux, 1 000 à 5 000 employés, Danemark

Ce rapport examine quatre points essentiels soulevés par les participants au sujet des stratégies réseau :

- **Point n°1 :** La planification des réseaux doit anticiper les besoins et ne pas se contenter de réagir aux cas d'usage actuels.
- **Point n°2 :** Les entreprises leaders s'appuient sur une infrastructure diversifiée pour l'IA, et les stratégies réseau doivent refléter cette réalité.
- **Point n°3 :** La latence, la bande passante et la disponibilité déterminent les stratégies réseau des entreprises les plus en pointe sur l'IA.
- **Point n°4 :** La gestion des workloads IA bénéficie de la segmentation du réseau.

Point n°1 : La planification des réseaux doit anticiper les besoins et ne pas se contenter de réagir aux cas d'usage actuels

À mesure que les projets d'IA passent de la phase pilote à l'opérationnalisation à grande échelle, les exigences accrues en matière de transfert de données (pour l'inférence et, dans certains projets, pour un entraînement complémentaire des modèles), peuvent peser lourdement sur les ressources réseau. Les participants à l'étude conviennent du fait que, pour assurer une communication efficace entre les nœuds, une infrastructure réseau solide est nécessaire. Nombreux sont ceux qui remarquent que cette problématique n'a pas été prise en compte lors de l'expérimentation initiale de l'IA. Dans un univers de l'IA générative où les projets sont soumis à un calendrier ambitieux, beaucoup d'entreprises se sont lancées sans réellement anticiper leurs besoins futurs en matière de réseau. La feuille de route IA de nombreuses entreprises devrait encore accentuer la pression sur les ressources.

« Lorsque vous commencez à automatiser les processus et à recourir plus souvent au LLM, vos latences, vos équipements, vos GPU... tout augmente. Je pense qu'au cours des prochaines années, la majorité des IA seront agentiques, ce qui signifie qu'il y aura plus de discussions, plus d'allers-retours entre le modèle et le workflow. »

Data scientist senior, GenAI
Services financiers, >100 000 employés, États-Unis

Les répondants d'Amérique du Nord considèrent généralement que leurs processus de déploiement de projets d'IA sont plus avancés que ceux des autres régions. Sur une échelle de 1 à 5, les entreprises nord-américaines ont attribué une note moyenne de 3,0 à la maturité de leurs processus de déploiement, devançant les entreprises européennes (2,1) et celles de la région Asie-Pacifique (2,2). Le défi que représente l'amélioration de la planification réseau peut être ressenti plus vivement par les entreprises moins avancées au niveau des processus. Toutefois, même aux États-Unis, aucune des personnes interrogées n'a estimé que les processus de son entreprise étaient aussi avancés qu'ils ne pourraient l'être.

Les participants à l'étude reconnaissent que la prochaine vague d'investissements dans l'IA augmentera considérablement la pression sur les ressources réseau. Cette prise de conscience découle de la constatation que de nombreuses applications d'IA émergentes exigeront soit une latence ultra-faible, soit un déploiement en périphérie (Edge Computing).

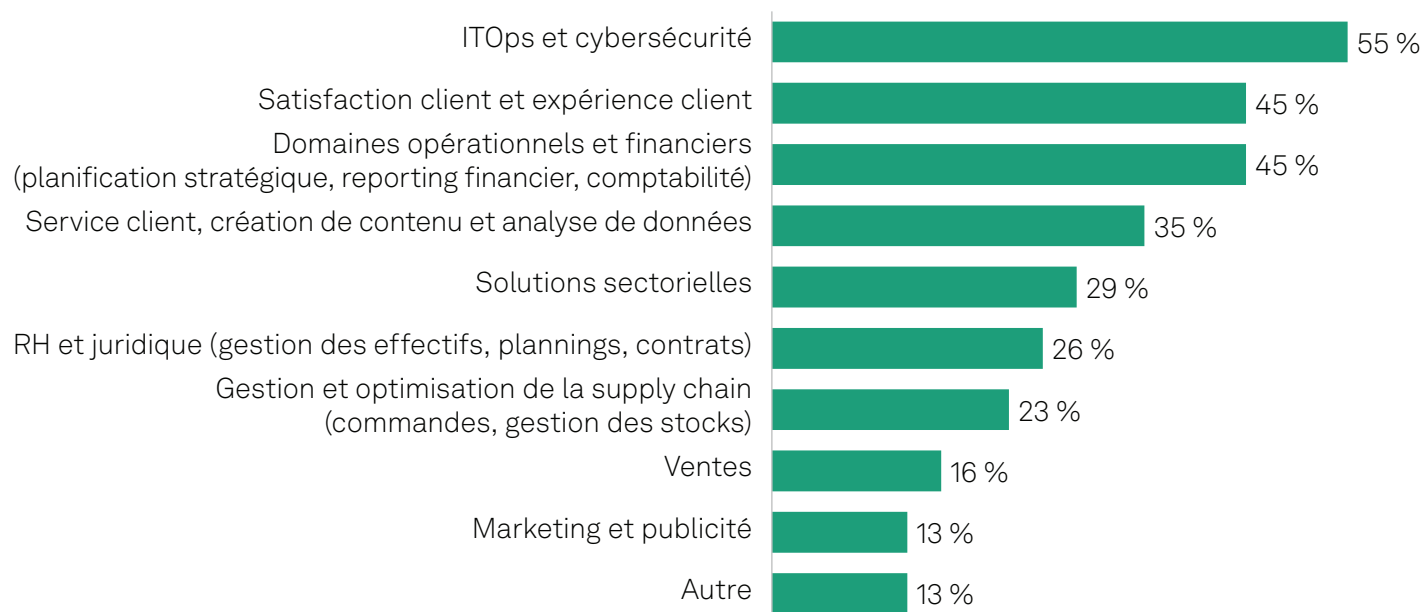
« Nos projets d'IA dans les domaines de l'analyse des données de vente en temps réel, de la S&OP [planification des ventes et des opérations] et des entrepôts devraient être achevés d'ici 6 à 12 mois. L'expérience que nous en avons tirée nous permettra d'orienter nos choix futurs entre cloud et on-prem... L'Edge Computing sera certainement pris en considération lorsque nous passerons à l'optimisation de l'IA en production. »

Responsable de la cybersécurité et des infrastructures

Fabrication de machines, 1 000 à 5 000 employés, Allemagne

Interrogés sur les applications avancées de l'IA, de nombreux participants ont identifié l'ITOps et la cybersécurité, la satisfaction et l'expérience client, ainsi que l'opérationnel et le financier comme des domaines d'investissement clés. Dans la plupart de ces cas d'usage, les applications d'IA risquent de repousser les limites de l'infrastructure existante. Par exemple, une interface conversationnelle conçue pour améliorer l'expérience client peut exiger des niveaux de réactivité plus élevés qu'une application interne. Un participant à l'étude a pris pour exemple le temps de traitement de près de cinq secondes d'un service d'IA de conversation vocale qu'il souhaitait implémenter, ce qui l'a contraint à explorer de nouvelles configurations et méthodes d'entraînement du modèle pour améliorer les temps de réponse.

Figure 2 : Domaines fonctionnels présentant les cas d'usage les plus avancés en matière d'IA/ML



Q. Parmi les domaines fonctionnels ci-dessous, veuillez indiquer les trois domaines qui mobilisent vos cas d'usage IA/ML les plus avancés ?

Base : Participants au groupe de discussion (n = 30).

Source : Étude à grande échelle réalisée pour Verizon par S&P Global Market Intelligence 451 Research AI.

Alors que certains répondants se sont concentrés sur des cas d'usage génériques, force est de constater que les applications d'IA spécifiques à un secteur ont été les plus difficiles à implémenter. C'est particulièrement vrai pour les répondants travaillant dans l'industrie, l'énergie, la santé et les sciences de la vie, où des cahiers des charges très stricts dictent l'architecture des infrastructures d'IA.

« Pour bien fonctionner, les IA utilisées pour le triage de patients souffrant de septicémie ou d'un AVC ont besoin de puissance de calcul en périphérie. Les images pour le calcul GPU localisé afin de traiter de vastes datasets d'imagerie hors cloud, mais aussi pour l'écoute et la transcription des échanges dans les salles d'examens médicaux. »

Directeur informatique senior
Santé/sciences de la vie, États-Unis

« Cas d'usage exigeant une faible latence pour l'intégration de l'IoT à des fins de contrôle qualité en temps réel. Les systèmes de contrôle qualité utilisent la vision par IA... Projets d'optimisation des transports en temps réel impliquant l'intégration de multiples sources de données. »

Directeur informatique
Industrie automobile, >50 000 employés, États-Unis

Figure 3 : Domaines d'application de l'IA dans l'industrie, la distribution d'énergie, la santé et les sciences de la vie

Industrie	Distribution d'énergie	Santé/sciences de la vie
- IoT, y compris la surveillance des produits et la maintenance prédictive - Analyse de la voix du client (VoC) - Systèmes de vision en temps réel pour le contrôle qualité et la production - Systèmes de transport en temps réel - Analyse des ventes en temps réel/ planification des ventes et des opérations - Systèmes d'entreposage - Optimisation des processus de fabrication/OEE - Innovation produit	- IoT sur les systèmes de contrôle et d'acquisition de données (SCADA) - Détection d'anomalies dans les conduites d'eau - Flux de données sur les stations d'épuration	- Triage médical - Traitement d'images - Transcription médicale - Surveillance des dispositifs médicaux - Suivi du comportement des patients - Évaluation des besoins en soins - Différents systèmes de vision - Monitoring des patients à distance - Suivi des traitements - Pharmacovigilance - Robot aide-soignant - Systèmes de réalité augmentée (p. ex. RA, RV) - Analyse des données de recherche et de développement

Source : Étude à grande échelle réalisée pour Verizon par S&P Global Market Intelligence 451 Research AI.

Les participants à l'étude ont insisté à plusieurs reprises sur les avantages de la planification, et nombre d'entre eux auraient souhaité développer plus tôt une vision plus orientée vers l'avenir. Une recommandation commune est de veiller à ce que les entreprises tiennent compte des questions de déploiement et de scalabilité au cours des phases de pilotes et d'entraînement de l'IA.

« Se concentrer... sur le déploiement du modèle autant que sur l'entraînement du modèle. »

Responsable architecture/cybersécurité
Secteur bancaire, 20 000 à 50 000 employés, Singapour

« Il est essentiel de planifier en tenant compte des opérations et de clarifier les exigences non fonctionnelles de l'environnement de production (performances, disponibilité et sécurité) avant de commencer le PoC. »

Directeur de la technologie
Services professionnels, >1 000 à 5 000 employés, Japon

« Dimensionnez vos besoins en GPU/CPU/stockage/mémoire pour les 18 mois à venir. Un débit de 10 Gbit/s devrait être la vitesse minimale du réseau, de bout en bout. »

Vice-président de l'architecture et de l'infrastructure IT
Santé, 20 000 à 50 000 employés, États-Unis

De nombreux participants à l'étude ont constaté des retards ou des abandons de projets en raison d'une planification initiale inadéquate des architectures réseau. Plusieurs ont indiqué que la vétusté de leur infrastructure technique et l'incapacité à intégrer le progrès technologique limitaient considérablement les possibilités de déploiement de l'IA et posaient de sérieux problèmes pour le transfert de données. Ce manque d'anticipation a souvent engendré des problèmes lors de l'opérationnalisation de l'IA, une phase pendant laquelle toute mise à niveau de l'infrastructure s'avère plus coûteuse, plus lente et plus difficile sur le plan technique. Les goulets d'étranglement du réseau doivent être anticipés et pris en compte dès les premières étapes de la planification.

« Les temps de réponse, qui ne posaient aucun problème dans l'environnement de développement, ont fortement ralenti (de plusieurs dizaines de minutes) dans l'environnement de production, et ce en raison d'un important volume de données transitant par le WAN et qui a saturé la bande passante disponible. »

Directeur de la stratégie IT

Fabricant d'électronique, >50 000 employés, Japon

« À un moment donné, nous avons eu un débordement de données en raison d'une mauvaise planification. Nous avons accumulé des quantités massives de données et avons dû résoudre les problèmes un par un, agir en mode pompier, ce que nous regrettons aujourd'hui. La scalabilité est importante pour éviter les débordements de données. »

Responsable des systèmes IT/R&D

Industrie, 20 000 à 50 000 employés, Japon

« Prévoyez plus que ce dont vous pensez avoir besoin. S'engager dans un processus IA/ML, puis s'arrêter pour mettre à niveau ou ajouter des équipements, finira par coûter plus cher que de suracheter au départ. »

Directeur de l'infrastructure d'ingénierie

Services financiers, >50 000 employés, États-Unis

Point n°2 : Les entreprises leaders s'appuient sur une infrastructure diversifiée pour l'IA, et les stratégies réseau doivent refléter cette réalité

Si de nombreuses entreprises expérimentent l'IA générative dans le cloud public, leurs workloads d'IA tendent ensuite à s'étaler sur différentes infrastructures au fur et à mesure qu'elles progressent dans leur démarche. La préparation des données, l'entraînement et l'inférence sont réparties sur différents sites, de la périphérie (edge) jusqu'au cœur du réseau, en passant par la pré-périphérie (near-edge). Cette complexité se reflète dans l'enquête de 451 Research Voice of the Enterprise: AI & Machine Learning, Infrastructure 2024, dans laquelle les répondants ont déclaré répartir l'inférence sur un large éventail d'infrastructures, y compris des hyperscalers du cloud public (61 %), des équipements informatiques (46 %), l'infrastructure de l'opérateur réseau (42 %), des data centers de proximité (41 %), des systèmes autonomes (39 %) et des clouds spécialisés dans l'IA (32 %).

Cette complexité architecturale est importante car, bien que certains sondés aient opté pour une stratégie réseau simple et centrée sur un seul CSP pour leurs projets d'IA, il s'agissait plutôt de l'exception que de la règle. Investissements existants, maintien de la confidentialité, impératifs de latence et de coût : pour toutes ces raisons, les entreprises ont élargi le champ d'action de leurs projets d'IA et introduit de nouveaux cas d'usage, tandis que se développaient leurs configurations architecturales en parallèle. De nombreux participants dotés d'une solide expérience en la matière ont indiqué que le cloud n'était pas toujours la réponse à leurs besoins en matière de workloads d'IA.

« En fonction du workload et du cas d'usage, nous choisissons la meilleure approche possible : cloud ou on-prem. Les considérations de coût sont un facteur déterminant dans notre prise de décision. »

Responsable de la cybersécurité et des infrastructures

Constructeur de machines, 1 000 à 5 000 employés, Allemagne

« L'Edge Computing permet de réduire considérablement les coûts d'entrée et de sortie du cloud, ainsi que les temps de latence. »

Responsable de l'infrastructure IT et de la sécurité

Retail, 1 000 à 5 000 employés, Allemagne

La majorité (54 %) des personnes interrogées dans les entreprises de plus de [1]20 000 salariées optent pour des environnements hybrides composés d'un mélange d'infrastructures on-prem et cloud. Un seul participant au groupe de discussion a décrit son approche actuelle comme étant principalement on-prem. Cette personne travaillant dans le secteur des services financiers a déclaré que cette stratégie était nécessaire pour « se conformer à nos nombreuses certifications qui nous permettent d'opérer ».

« Nous adoptons une approche hybride : l'entraînement de l'IA est effectué dans le cloud, tandis que l'inférence en temps réel au point de service est effectuée on-prem. Cela nous permet d'équilibrer efficacement les coûts et les exigences en matière de niveau de service... Pour le traitement des données sensibles (telles que les informations confidentielles ou personnelles qui ne peuvent pas être envoyées à l'extérieur), nous nous appuyons souvent sur une infrastructure on-prem pour garantir la sécurité des données. »

Responsable du laboratoire de technologies avancées

Retail, >50 000 employés, Japon

Les entreprises qui n'ont pas tenu compte de la connectivité entre les environnements, de la complexité du routage, des besoins du réseau interne et des questions de bande passante se sont retrouvées en difficulté lorsque la charge a augmenté et que les volumes de données se sont accrus. Interrogées sur leurs principales préoccupations et difficultés, les entreprises ont cité à maintes reprises la nécessité de déplacer les données d'un environnement à l'autre.

« Notre plus grande préoccupation actuellement est le coût du transfert de grands datasets de l'on-prem vers Azure et d'en baisser les coûts. »

Responsable infrastructure senior

Santé, 20 000 à 50 000 employés, États-Unis

« Nous sommes confrontés à des ralentissements de l'entraînement lorsque nous déplaçons de grands volumes de données d'une zone à l'autre, et l'inférence sur des sites distants est un problème en raison de l'instabilité ou de la limitation des options de connexion. »

Directeur informatique senior

Santé/sciences de la vie, 10 000 à 20 000 employés, États-Unis

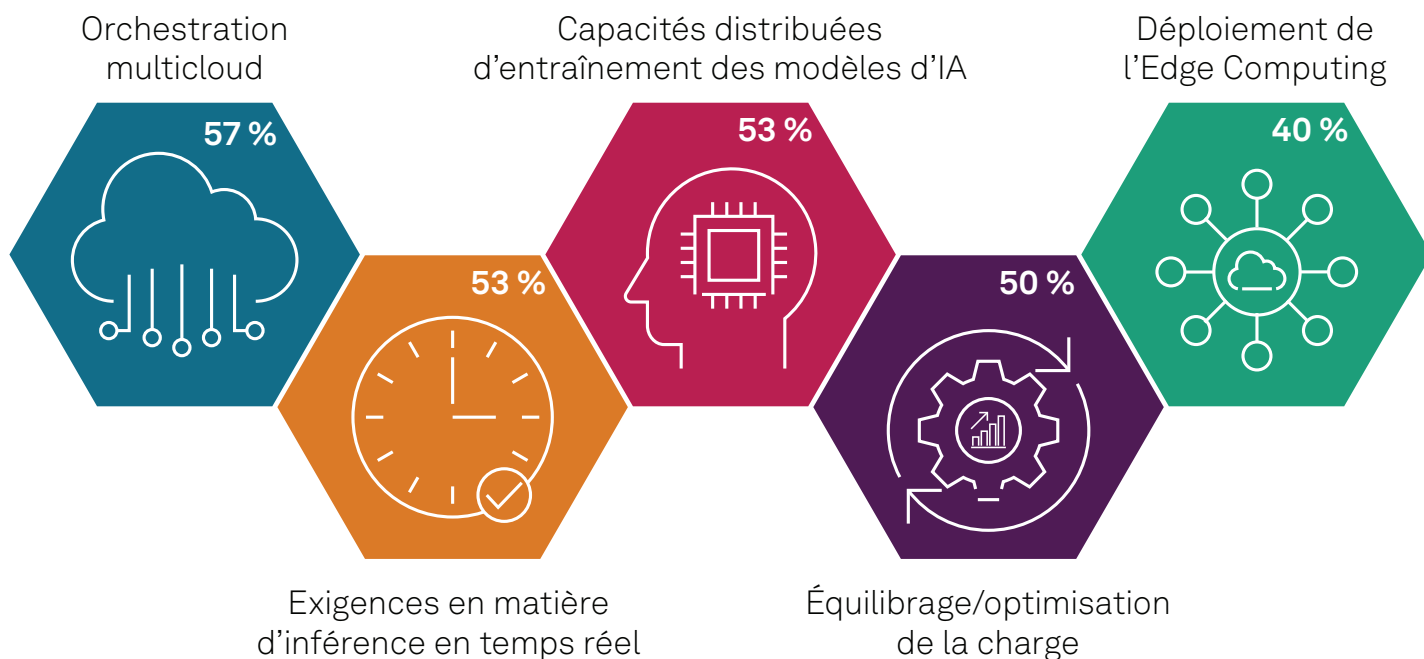
« Certaines données sont échangées à l'échelle mondiale via le réseau privé de notre entreprise. Avec l'utilisation croissante de l'IA, il y a un fort besoin d'augmenter de manière significative la bande passante. »

Directeur de la stratégie IT

Fabricant d'électronique, >50 000 employés, Japon

Lorsqu'on les interroge sur les besoins d'infrastructure les plus critiques pour leurs projets d'IA, les sondés citent d'abord l'orchestration multicloud de l'IA distribuée, suivie de près par des capacités distribuées d'entraînement des modèles et l'inférence en temps réel. Les répondants ont également souligné la nécessité de prendre en compte la localisation des données, l'efficacité de leur transfert, le maintien de versions cohérentes des modèles et l'orchestration.

Figure 4 : L'orchestration multicloud émerge comme un besoin d'infrastructure essentiel



Q. Quels sont les besoins d'infrastructure émergents les plus importants pour vos projets d'IA/ML à l'avenir ?

Base : Participants au groupe de discussion (n = 30).

Étude à grande échelle réalisée pour Verizon par S&P Global Market Intelligence 451 Research AI.

Plusieurs responsables IA ont souligné l'importance d'établir un cadre plus clair pour déterminer le lieu d'exécution de certains workloads spécifiques. Un cadre-dirigeant a souligné la nécessité d'impliquer et d'informer les développeurs d'applications d'IA et les data scientists à ce sujet, et ce afin de garantir une bonne implémentation du cadre en question.

« Nous essayons d'avoir une stratégie... nous essayons de construire un processus... pour guider les différentes équipes de data scientists, pour formuler des recommandations sur les cas d'usage les mieux adaptés à l'on-prem ou au cloud. Ils ont besoin de notre certification pour déployer... quand utiliser le traitement par batch, l'inférence en temps réel et d'autres choses de ce genre. »

Directeur senior, produits et plateformes d'IA

Services financiers, 20 000 à 50 000 employés, États-Unis

Point n°3 : La latence, la bande passante et la disponibilité déterminent les stratégies réseau des entreprises les plus en pointe sur l'IA

À mesure que les entreprises développent leurs applications d'IA, le besoin d'une infrastructure à faible latence et à large bande passante apparaît comme un fondement essentiel de la planification du réseau. De nombreux participants ont souligné la nécessité d'élaborer des stratégies réseau visant à assurer un transfert efficace de grands volumes de données et de faciliter le traitement en temps réel. En matière d'architecture, les conseils s'articulent généralement autour de trois facteurs : la disponibilité, la latence et la bande passante.

« Visez une large bande passante et une faible latence. »

Responsable des services d'infrastructure
Secteur bancaire, 20 000 à 50 000 employés, Australie

Les entreprises qui ont mené à bien leurs projets d'IA à grande échelle ont déjà investi dans des pratiques rigoureuses pour assurer la disponibilité du réseau. Lorsqu'on leur a demandé d'évaluer leur efficacité sur une échelle de 1 à 5, 92 % d'entre eux se sont attribués une note de 4 ou 5, ce qui est supérieur à toutes les autres pratiques évaluées en matière d'infrastructure. Malgré un tel niveau d'avancement, ce domaine est jugé suffisamment critique pour que la plupart d'entre eux identifient un besoin d'investissement continu.

« La disponibilité du réseau garantit un flux de données temps réel précises qui fourniront les analyses nécessaires à la prise de décisions. »

Directeur de l'infrastructure et du cloud
Gaming, 1 000 à 5 000 employés, Suède

« Établir des connexions à haut débit et à faible latence en utilisant des lignes dédiées pour se connecter à des systèmes on-prem et à d'autres environnements cloud. »

Directeur de la stratégie IT
Fabricant d'électronique, >50 000 employés, Japon

« Concevez une infrastructure évolutive avec des réseaux à haut débit et à faible latence. Optimisez les flux de données grâce au cache et au stockage distribué. »

Directeur de l'ingénierie cloud
Assurance, 20 000 à 50 000 employés, États-Unis

Il est crucial de mettre l'accent sur la latence, la disponibilité et la bande passante, car les personnes interrogées ont cité de nombreux cas où les limites du réseau ont fait dérailler des projets ou, du moins, exercent une forte contrainte sur leurs initiatives d'IA.

« Le réseau entrave notre capacité à déployer davantage de fonctionnalités en raison de problèmes de latence et de la taille des connexions. »

Directeur senior des opérations et de la stratégie IT
Fabricant de produits chimiques, 1 000 à 5 000 employés, États-Unis

« Parfois, la segmentation existante et l'absence de contrôle des flux de données en temps réel empêchent la montée en charge de notre application d'IA temps réel, laquelle s'étend sur plusieurs pays. »

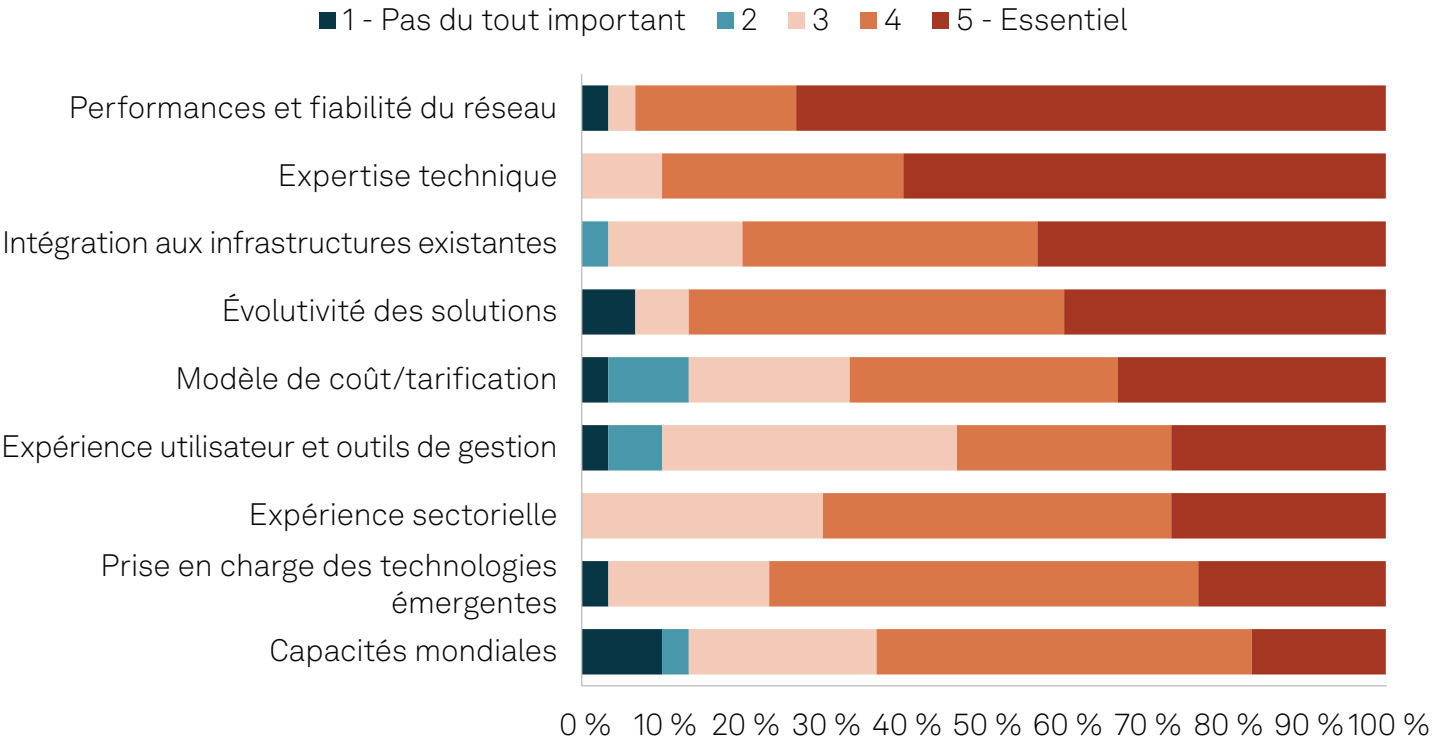
Responsable architecture/cybersécurité
Secteur bancaire, 20 000 à 50 000 employés, Singapour

« L'optimisation du transfert de données (en particulier l'amélioration de la bande passante et de la latence) est un domaine clé à prendre en compte... La raison en est que sur les workloads AI/ML, de grands volumes de données sont transférés vers le cloud, et la bande passante peut facilement devenir un goulet d'étranglement. »

Directeur de la technologie
Services professionnels, >1 000 à 5 000 employés, Japon

Les participants ont indiqué que la performance et la fiabilité du réseau étaient les facteurs les plus importants à prendre en compte lors de l'évaluation des partenaires technologiques pour le réseau et la sécurité. Les participants ont insisté sur ce domaine plus encore que sur le coût, l'expérience du secteur et l'expérience utilisateur (voir figure 5). Cela renforce une thématique centrale de l'étude : une stratégie d'infrastructure d'IA efficace doit être axée sur la performance réseau.

Figure 5 : Principales capacités des partenaires technologiques pour l'infrastructure réseau/de sécurité réseau



Q. Lorsque vous évaluez d'éventuels partenaires technologiques pour l'infrastructure réseau/de sécurité réseau qui soutient vos initiatives d'IA/ML, veuillez indiquer l'importance des éléments suivants (1 = Pas du tout important ; 5 = Essentiel).
Base : Participants au groupe de discussion (n = 30).
Source : Étude à grande échelle réalisée pour Verizon par S&P Global Market Intelligence 451 Research AI.

Point n°4 : La gestion des workloads IA bénéficie de la segmentation du réseau

Les participants à l'étude ont souligné l'importance d'établir des sous-réseaux dédiés spécifiquement au matériel compatible avec l'IA. Interrogés sur les bonnes pratiques infrastructurelles pour les workloads d'IA, de nombreux répondants ont cité la segmentation du réseau, c'est-à-dire la création de segments réseau isolés et soumis à des politiques de routage adaptées aux workloads d'IA.

« Séparation en fonction de l'environnement et du cycle de vie IA/ML, c'est-à-dire entre l'entraînement et l'opérationnel. »

Vice-président de l'architecture

Services professionnels, 1 000 à 5 000 employés, Royaume-Uni

« Créez une architecture en couches de zones. »

Directeur mondial de l'infrastructure et des opérations

Industrie automobile/aérospatiale, 10 000 à 20 000 employés, Allemagne

« Pour l'on-prem, nous mettons en place une extension du réseau et créons un sous-réseau distinct pour le matériel compatible IA/ML, avec des routages spécifiques. »

Responsable d'hébergement et de services cloud

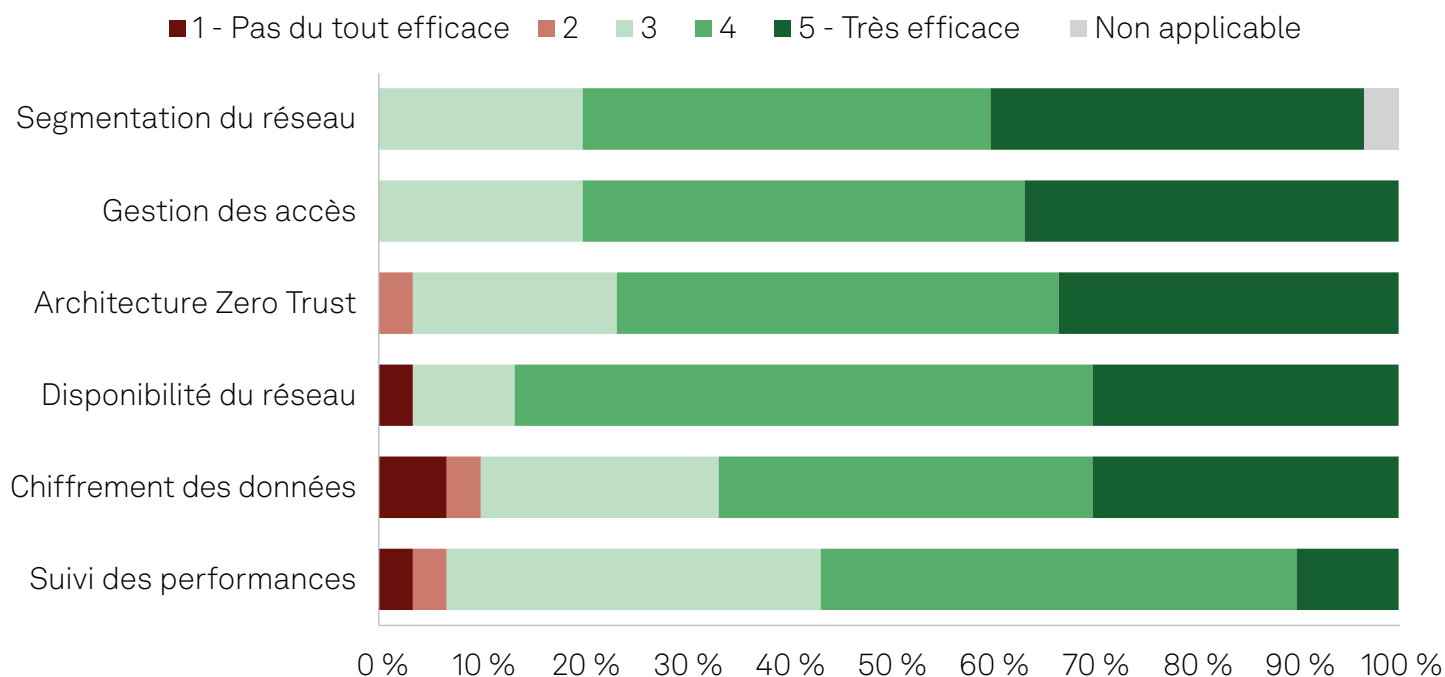
Services de distribution d'eau, 1 000 à 5 000 employés, Australie

Deux facteurs principaux sont à l'origine de cette tendance à la segmentation du réseau. Tout d'abord, les participants à l'enquête ont noté que les applications d'IA à fort impact génèrent généralement un trafic réseau important. En isolant ce trafic, les entreprises peuvent exercer un meilleur contrôle, éviter la congestion du réseau et s'assurer que les processus d'IA/ML n'interfèrent pas avec d'autres opérations. Les performances des clusters de GPU utilisés pour l'entraînement de l'IA dépendent fortement du débit et de la latence du réseau. En excluant le trafic non lié à l'IA, vous optimisez les performances du réseau.

Ensuite, la segmentation du réseau peut limiter l'exposition des workloads d'IA, ajoutant des couches de sécurité spécifiquement configurées pour l'IA et garantant la protection des données sensibles. Les modèles d'IA et leur concentration d'informations critiques en font des cibles privilégiées pour les cyberattaquants. L'isolation par couches qui procure la segmentation est un avantage clé en matière de sécurité. Ce sujet sera examiné plus en détail dans notre prochain rapport sur la sécurité.

Les entreprises qui ont mené à bien des projets d'IA à grande échelle ont investi massivement dans la segmentation du réseau (voir figure 6).

Figure 6 : Efficacité des pratiques d'infrastructure pour les workloads d'IA



Q. Veuillez évaluer l'efficacité des pratiques suivantes pour vos workloads IA/ML (1 = pas du tout efficace, 5 = très efficace).

Base : Participants au groupe de discussion (n = 30).

Source : Étude à grande échelle réalisée pour Verizon par S&P Global Market Intelligence 451 Research AI.

« Nous disposons d'une architecture réseau Zero Trust segmentée, avec un sous-réseau pour l'entraînement, l'inférence et les pipelines de données. »

Responsable architecture/cybersécurité

Secteur bancaire, 20 000 à 50 000 employés, Singapour

« La segmentation du réseau est appliquée à l'échelle de la banque... Elle est gérée par une équipe centralisée qui suit des processus SRE très robustes. »

Responsable architecture/cybersécurité

Secteur bancaire, 20 000 à 50 000 employés, Singapour

En matière de segmentation des réseaux, le niveau de maturité n'est pas toujours en corrélation avec les règles de confidentialité des données propres au secteur dans lequel l'entreprise évolue. En moyenne, les répondants du secteur de la santé et des sciences de la vie font état de stratégies de segmentation réseau moins matures que celles d'autres secteurs. Plusieurs répondants travaillant dans les services financiers ont également indiqué qu'il existait encore une marge de progression.

L'accent mis sur la segmentation des réseaux est particulièrement remarquable en raison de la quantité des problèmes qu'une telle opération peut soulever, à commencer par le coût. Lorsqu'on a demandé aux participants au groupe de discussion quels étaient les trois principaux obstacles à l'implémentation de l'IA, 58 % ont cité des problèmes de coûts, un chiffre qui augmente si l'on considère la question du retour sur investissement. En outre, l'étude montre que les équipes d'infrastructure sont souvent déjà à bout de souffle. Ainsi, le fardeau que représentent le déploiement et la gestion de la segmentation pourrait encore accentuer la pression sur les équipes réseau et sécurité. L'automatisation des réseaux peut apporter des éléments de réponse, sauf que les entreprises ont toujours eu du mal à la maîtriser.

De même, de nombreux participants voient dans la complexité de l'architecture existante un obstacle à l'implémentation de l'IA. Cette complexité a des répercussions sur les exigences d'intégration que toute stratégie de segmentation du réseau doit prendre en compte. Elle peut également avoir une incidence sur la capacité à maintenir la continuité des activités au moment de la segmentation du réseau. Il est donc particulièrement intéressant de noter que, malgré tous ces problèmes, un large échantillon de participants à l'étude a souligné l'importance de la segmentation.

Implications

Peu de managers et dirigeants s'interrogent sur la valeur de l'IA pour leur entreprise. Pour eux, la question qui domine est celle de son exploitation la plus efficace possible. Dans l'étude de 451 Research Voice of the Enterprise: IA & Machine Learning, Use Cases 2025, seulement 5 % des répondants déclarent que l'IA générative n'est pas utilisée dans leur organisation, tandis que 11 % rapportent qu'elle n'était utilisée que de manière informelle, sur une base individuelle. De même, seulement 27 % ont déployé l'IA générative dans l'ensemble de leur organisation, la majorité des organisations étant encore au stade de l'expérimentation technologique ou d'une opérationnalisation de déploiements à plus petite échelle.

Qu'est-ce qui empêche de nombreuses entreprises de développer leurs projets d'IA ? Les participants qui sont parvenus à déployer l'IA à grande échelle considèrent l'infrastructure réseau comme un potentiel frein majeur dans la transition. Lorsque les entreprises élaborent une stratégie réseau visant à soutenir un déploiement opérationnel à grande échelle, leur réflexion doit s'articuler autour de quatre points clés :

Point n°1 : La planification des réseaux doit anticiper les besoins et ne pas se contenter de réagir aux cas d'usage actuels.

Les dirigeants doivent éviter le piège de la réactivité et, au contraire, se concentrer sur une planification proactive du réseau. Plutôt que de se concentrer sur les besoins immédiats des pilotes et de l'entraînement de modèles, les entreprises devraient anticiper les exigences des projets d'IA en production, en cherchant à estimer à l'avance les ressources nécessaires. Cela pourra nécessiter une plus grande clarification des exigences non fonctionnelles telles que la performance, la disponibilité et la sécurité, et ce avant même de commencer les preuves de concept (PoC). Il s'agira également de prévoir des capacités excédentaires pour faire face à la croissance et éviter les retards dans les projets.

Point n°2 : Les entreprises leaders s'appuient sur une infrastructure diversifiée pour l'IA, et les stratégies réseau doivent refléter cette réalité.

Si les premières expérimentations peuvent avoir lieu dans le cloud, la mise en production des projets d'IA nécessitera souvent des options d'infrastructure différentes. La stratégie réseau doit ainsi être suffisamment souple pour prévoir l'extension de la charge à différents environnements, de la périphérie jusqu'au cœur, et d'anticiper le besoin de connectivité entre ces environnements. Les infrastructures actuelles sont généralement hybrides et les investissements réseau doivent en tenir compte. En établissant des cadres clairs pour l'affectation de cette charge, et en impliquant des développeurs d'applications d'IA et des data scientists dans le processus décisionnel, vous pourrez atténuer une partie de cette complexité architecturale.

Point n°3 : La latence, la bande passante et la disponibilité déterminent les stratégies réseau des entreprises les plus en pointe sur l'IA.

Pour faire face aux énormes volumes de données, ainsi qu'aux exigences de traitement en temps réel associées à l'IA, les décideurs doivent faire de la latence et de la bande passante les éléments fondamentaux de la planification de leur réseau. Il est conseillé de se concentrer sur le triptyque disponibilité, latence et bande passante lors de la conception des architectures réseau, en cherchant éventuellement à optimiser davantage les flux de données à l'aide du stockage distribué et de la mise en cache. Un investissement proactif et continu dans la disponibilité du réseau peut aider les entreprises à éviter tout déraillement de leurs projets.

Point n°4 : La gestion des workloads IA bénéficie de la segmentation du réseau.

D'où l'importance d'établir des segments réseau dédiés aux clusters d'IA et de prévoir une capacité adéquate pour le déplacement des données et des modèles à travers l'infrastructure on-prem et le cloud. Les entreprises peuvent également créer des segments de réseau isolés, avec des politiques de routage personnalisées, pour gérer le fort trafic réseau et éviter les congestions, en tenant compte des différents environnements et des étapes du cycle de vie de l'IA/ML, notamment l'entraînement des modèles par rapport à l'opérationnel.

Ce rapport est le premier d'une série en trois parties sur les retours d'expérience de leaders de l'IA en matière de lancement, d'opérationnalisation et de rentabilisation de projets d'IA. Le deuxième rapport de cette série portera sur la sécurité, et le troisième dressera un bilan des bonnes pratiques globales.



À propos de Verizon

Verizon Business a pour mission d'aider des entreprises comme la vôtre à réussir dans l'environnement numérique dynamique d'aujourd'hui. Nous fournissons des solutions réseau essentielles qui sous-tendent et optimisent les opérations métiers et permettent à des millions de personnes de vivre, de travailler et de se divertir chaque jour. Cette étude a été commissionnée afin de mieux comprendre les enjeux du déploiement de l'IA, de partager les expériences des pairs et de clarifier les choix stratégiques auxquels sont confrontés les responsables IT. Notre objectif est d'offrir une meilleure compréhension de ces dynamiques complexes, en particulier de l'importance critique de l'infrastructure réseau et sécurité. Nous pensons que ces informations peuvent aider votre entreprise à développer l'IA en toute confiance, à maximiser son potentiel et à garder une longueur d'avance dans cet environnement en constante évolution.

À propos de l'auteur



Alex Johnston

Chercheur-analyste senior

Alex Johnston est chercheur-analyste senior au sein de l'équipe 451 Research Data, AI & Analytics de S&P Global Market Intelligence. Il se concentre sur les technologies émergentes et sur la manière dont elles peuvent être appliquées dans des contextes business. Les principaux domaines couverts par M. Johnston sont l'intelligence artificielle, la technologie des grands livres distribués, le traitement des flux d'événements et les marketplaces de données. Il s'est récemment concentré sur le suivi du marché émergent de l'IA générative, sur l'évolution des cas d'usage de la blockchain et sur l'étude des architectures temps réel.

À propos de ce rapport

Un rapport de découverte est une étude basée sur des données d'enquête de recherche préliminaire, qui évalue la dynamique du marché d'un segment technologique clé de l'entreprise à travers l'expérience « sur le terrain » et les opinions de véritables praticiens, ce qu'ils font et pourquoi ils le font.

À propos de S&P Global Market Intelligence

Chez S&P Global Market Intelligence, nous comprenons l'importance de bénéficier d'informations précises, approfondies et pertinentes. Notre équipe d'experts fournit des informations inégalées et des solutions de pointe en matière de données et de technologie, en partenariat avec les clients pour élargir leur perspective, opérer en toute confiance et prendre des décisions avec conviction.

S&P Global Market Intelligence est une division de S&P Global (NYSE : SPGI). S&P Global est le premier fournisseur mondial de notations de crédit, d'indices de référence, d'analyses et de solutions de flux de travail sur les marchés mondiaux des capitaux, des matières premières et de l'automobile. Chacune de nos offres aide de nombreuses organisations parmi les plus importantes au monde à naviguer dans le paysage économique afin qu'elles puissent préparer l'avenir, dès aujourd'hui. Pour plus d'informations, consultez le site www.spglobal.com/marketintelligence.

CONTACTS

Amériques : +1 800 447 2273

Japon : +81 3 6262 1887

Asie-Pacifique : +60 4 291 3600

Europe, Moyen-Orient, Afrique : +44 (0) 134 432 8300

www.spglobal.com/marketintelligence

www.spglobal.com/en/entreprise/about/contact-us.html

Copyright © 2025 par S&P Global Market Intelligence, une division de S&P Global Inc. Tous droits réservés.

Ces documents ont été préparés uniquement à des fins d'information sur la base d'informations généralement disponibles au public et provenant de sources jugées fiables. Aucun contenu (y compris les données relatives aux indices, les notations, les analyses et données liées aux crédits, la recherche, le modèle, le logiciel ou toute autre application ou donnée de ceux-ci) ou toute partie de celui-ci (le « Contenu ») ne peut être modifié, rétro conçu, reproduit ou distribué sous quelque forme que ce soit et par quelque moyen que ce soit, ni stocké dans une base de données ou un système d'extraction, sans l'autorisation écrite préalable de S&P Global Market Intelligence ou de ses sociétés affiliées (collectivement, « S&P Global »). Le contenu ne doit pas être utilisé à des fins illégales ou non autorisées. S&P Global et les fournisseurs tiers (collectivement « les Parties de S&P Global ») ne garantissent pas l'exactitude, l'exhaustivité, l'actualité ou la disponibilité du Contenu. Les Parties de S&P Global ne sont pas responsables des erreurs ou omissions, quelle qu'en soit la cause, des résultats obtenus par l'utilisation du Contenu. LE CONTENU EST FOURNI « EN L'ÉTAT ». LES PARTIES DE S&P GLOBAL REJETTENT TOUTE GARANTIE EXPRESSE OU IMPLICITE, Y COMPRIS, MAIS SANS S'Y LIMITER, TOUTE GARANTIE DE QUALITÉ MARCHANDE OU D'ADÉQUATION À UN USAGE PARTICULIER, D'ABSENCE DE BOGUES, D'ERREURS OU DE DÉFAUTS LOGICIELS, DE FONCTIONNEMENT ININTERROMPU DU CONTENU OU DE FONCTIONNEMENT DU CONTENU AVEC TOUTE CONFIGURATION LOGICIELLE OU MATÉRIELLE. Les Parties de S&P Global ne peuvent en aucun cas être tenues responsables envers une partie pour tous les dommages directs, indirects, accessoires, exemplaires, compensatoires, punitifs, spéciaux ou consécutifs, les coûts, les dépenses, les frais juridiques ou les pertes (y compris, mais sans s'y limiter, les pertes de revenus ou de profits, les coûts d'opportunité ou les pertes causées par la négligence) en rapport avec l'utilisation du Contenu, même si elles ont été informées de l'éventualité de tels dommages.

Les opinions, cotations et analyses de crédit et autres de S&P Global Market Intelligence sont des déclarations d'opinion à la date à laquelle elles sont exprimées et non des déclarations de fait ou des recommandations d'acheter, de détenir ou de vendre des titres ou de prendre des décisions d'investissement, et ne traitent pas de l'adéquation d'un titre. S&P Global Market Intelligence peut fournir des données sur les indices. L'investissement direct dans un indice n'est pas possible. L'exposition à une classe d'actifs représentée par un indice est disponible par le biais d'instruments d'investissement basés sur cet index. S&P Global Market Intelligence n'assume aucune obligation de mettre à jour le Contenu après sa publication sous quelque forme ou format que ce soit. Le Contenu ne doit pas être invoqué et ne remplace pas les compétences, le discernement et l'expérience de l'utilisateur, de sa direction, de ses employés, de ses conseillers et/ou de ses clients lorsqu'ils prennent des décisions d'investissement ou d'autres décisions commerciales. S&P Global maintient certaines activités de ses divisions séparées les unes des autres afin de préserver l'indépendance et l'objectivité de leurs activités respectives. Par conséquent, certaines divisions de S&P Global peuvent disposer d'informations qui ne sont pas accessibles à d'autres divisions de S&P Global. S&P Global a mis en place des politiques et des procédures visant à préserver la confidentialité de certaines informations non publiques reçues dans le cadre de chaque processus analytique.

S&P Global peut recevoir une rémunération pour ses notations et certaines analyses, normalement de la part d'émetteurs ou de souscripteurs de titres ou de débiteurs. S&P Global se réserve le droit de diffuser ses opinions et analyses. Les notations et analyses publiques de S&P Global sont disponibles sur ses sites web, www.standardandpoors.com (gratuit) et www.ratingsdirect.com (abonnement), et peuvent être diffusées par d'autres moyens, notamment par le biais des publications de S&P Global et de redistributeurs tiers. Des informations supplémentaires sur nos frais de notation sont disponibles à l'adresse suivante : www.standardandpoors.com/usratingsfees.