N°235 — Avril 2025

www.linformaticien.com

L'INFORMATICIEN

Ressources Humaines

« Nous recrutons beaucoup de développeurs et de Tech leads »

Rencontre avec Sylvie Verstraeten, DRH Docaposte

Réseau

Test Netgear Orbi

ESN

Nexpublica

Hardware

Xeon 6

Logiciel

Agentforce ouvre sa boutique

Retex

Audi virtualise sa production

DOSSIER

Observabilité

La performance par les données

L 14614 - 235 - F: 8,50 € - RD

VERTIV™

keep it humming™

Solution complète pour la révolution de l'IA

Alimentez et refroidissez l'IA de vos clients grâce à une solution complète unique.

L'IA est là et elle s'accompagne d'une demande d'alimentation et de refroidissement inédite. Dénouez les complexités grâce à Vertiv™ 360AI, des solutions complètes pour alimenter et refroidir de manière transparente les charges de travail d'IA et d'informatique accélérée.

En savoir plus :
vertiv.com/ai-hub-fr

2024 PALMARÈS L'INFORMATICIEN

Catégorie
Hardware Infrastructures
Critiques Data Center

www.linformaticien.com
L'INFORMATICIEN

RÉDACTION

88 boulevard de la Villette, 75019 Paris, France.
Tél. : +33 (0)1 74 70 16 30 — *contact@linformaticien.com*

RÉDACTION : Bertrand Garé (rédacteur en chef)
et Victor Miget (rédacteur en chef adjoint)
avec : François Cointe, Oscar Barthe, Patrick Brebion, Olivier Bouzereau, Vincent Bussière, Jérôme Cartegini, Michel Chotard, Alain Clapaud, Guillaume Renouard etThierry Thaureaux

SECRÉTAIRE DE RÉDACTION : Amélie Ermenault Martin

MAQUETTE ET RÉALISATION : Franck Soulier (chef de studio)

PUBLICITÉ
Antoine Foulon — *afoulon@linformaticien.com*

VENTE AU NUMÉRO

France métropolitaine 8,50 € TTC (TVA 5,5 %)

ABONNEMENTS

France métropolitaine 72 € TTC (TVA 5,5 %)
magazine + numérique

Toutes les offres :
www.linformaticien.com/abonnement

Pour toute commande d'abonnement d'entreprise ou d'administration avec règlement par mandat administratif, adressez votre bon de commande à :

L'Informaticien, service abonnements,
88 boulevard de la Villette, 75019 Paris, France.
ou à abonnements@linformaticien.com

IMPRESSION

Imprimé en France par Imprimerie Chirat (42)
Dépôt légal : 2ème trimestre 2025

L'INFORMATICIEN est publié par PC PRESSE, S. A. S.
au capital de 130 000 euros.
Siège social : 88 boulevard de la Villette, 75019 Paris, France.

ISSN 1637-5491

Une publication PCpresse,

FICADE

PRÉSIDENT, DIRECTEUR DE LA PUBLICATION :
Gaël Chervet

L'observabilité pour y voir clair !

Notre dossier mensuel fait le point sur les évolutions d'un concept apparu il y a déjà quelque temps : l'observabilité, le dernier épigone de ce qui fût la monitoring, la supervision, l'hyper-vision, petite fille de l'APM (Application Performance Monitoring). Elle est désormais l'arme ultime pour détecter et anticiper les problèmes et incidents dans les systèmes d'information, afin comme toujours d'améliorer la performance et la continuité des services offerts. Elle est aussi utile pour la cybersécurité et propose de voir l'ensemble de ce qui se passe dans les systèmes et applications pour y voir enfin clair avec des rapports lisibles permettant de prendre des actions rapides. Elle encourage une nouvelle évolution vers une informatique en plateforme et non un empilement d'outils de suivi et de rapports disparates pour avoir un point clair sur la situation de l'informatique de l'entreprise, du réseau, des logiciels utilisés. Le concept reste encore cependant assez confus du fait de la multitude de points couverts. Elle est donc devenue une sorte de mot-valise qui recouvre bien des aspects différents. Le dossier essaie donc d'y remédier autant que faire se peut.

Vous retrouverez bien sûr toutes nos rubriques avec le plein d'actualité aux côtés de ce dossier. Mais ce n'est pas tout. Le Top Tech est là ! Entreprises, ESN, il est temps de mettre vos projets en valeur, en présentant des dossiers dans les différentes catégories, afin que notre jury de personnalités représentatives puissent les classer. Résultats le 3 juillet prochain au Pavillon d'Armenonville pour les gagnants et les autres. Il faut rappeler que le dépôt des dossiers est gratuit. Vous trouverez toutes les informations nécessaires sur notre site ou directement à cette adresse : https://toptech.linformaticien.com

Bonne lecture, et souyez nombreux à poser des dossiers pour un des moments forts de l'année pour l'Informaticien. □

Bertrand Garé
Rédacteur en Chef

SMART IMPACT

THOMAS HUGUES

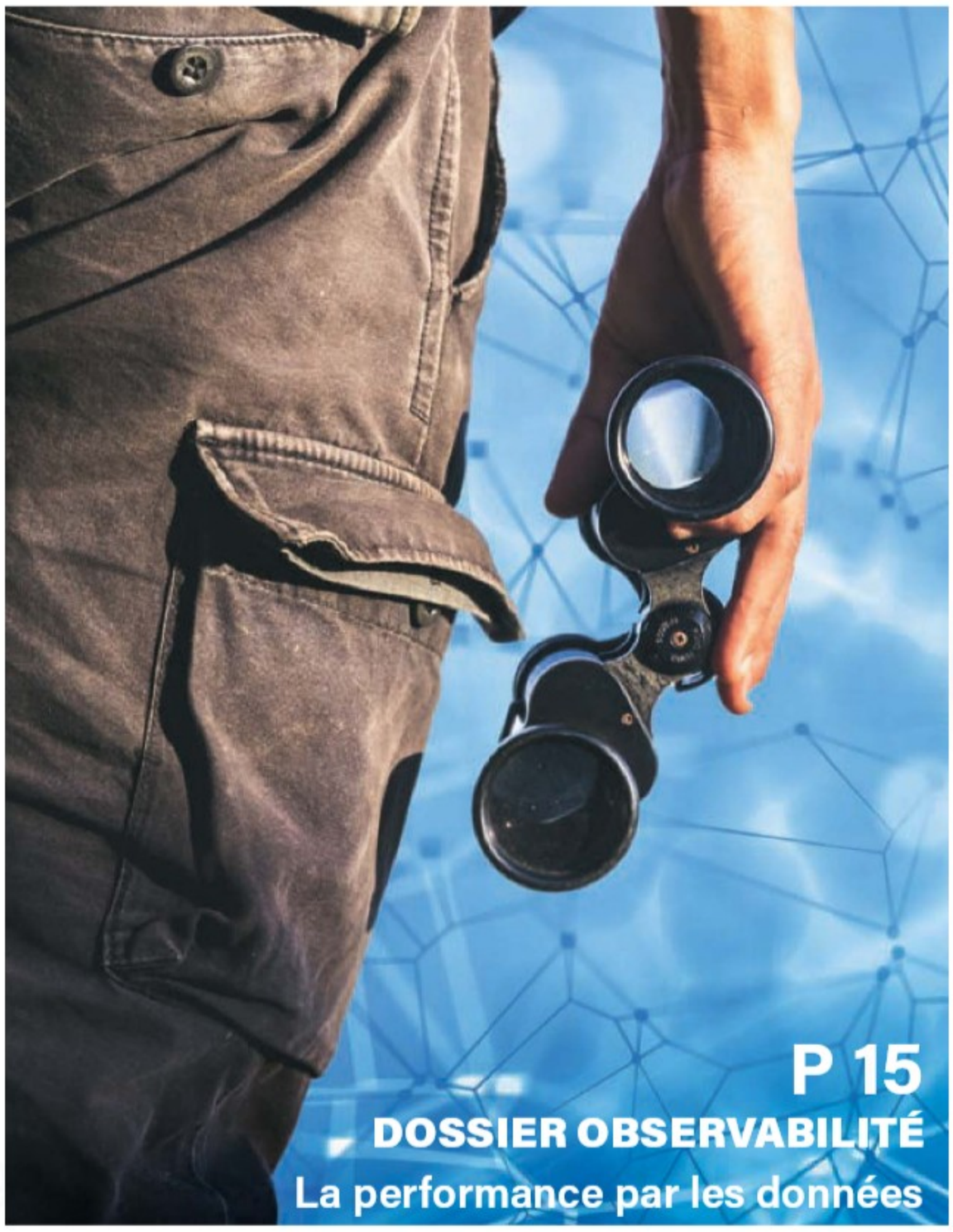
8H30 | 19H30

VOTRE ÉMISSION QUOTIDIENNE DÉDIÉE À LA RSE ET À LA TRANSITION ÉCOLOGIQUE DES ENTREPRISES

Orientée « solutions », l'émission SMART IMPACT animée par Thomas Hugues monte en puissance et vous propose désormais un rendez-vous quotidien. Chaque jour, retrouvez des témoignages d'entrepreneurs et d'experts autour de la transition écologique, de l'économie durable et des enjeux RSE.

N°230 orange™ | N°245 bouygues | N°349 free

B SMART 41. Change



P 15
DOSSIER OBSERVABILITÉ
La performance par les données

P 65
RH
Rencontre avec Sylvie Verstraeten, DRH Docaposte

P 35
RESEAUX
Routeur Netgear Orbi 970

Le Grand Saut : comment les DSI des ETI relèvent le défi du digital ?

La transformation numérique est une priorité pour les ETI françaises, et leurs DSI doivent innover tout en assurant la résilience et la sécurité des systèmes.

L'étude PAC « Le Grand Saut », menée en exclusivité pour Ready For IT, explore les ambitions et les obstacles rencontrés par ces décideurs.

Le numérique, un levier de compétitivité incontournable

Près de 80 % des DSI interrogés considèrent le digital comme le moteur central de leur stratégie de croissance. Comme en référence à un grand saut, l'étude met en évidence que la situation des DSI (spécifiquement en ETI) est en pleine évolution sur l'année 2025.

Et ce, à trois niveaux différents :

1. Un contexte économique et financier qui pousse nos entreprises à une performance globale de la part de toutes ses composantes, avec la DSI en première ligne.
2. L'aspect technologique n'y coupe pas : le déploiement massif des IA et les besoins toujours plus frénétiques de trouver LA nouvelle technologie pour aller encore plus vite, encore plus loin (vs. la concurrence).
3. Le point de vue réglementaire avec les nouvelles directives NIS 2, entre autres.

Face à ces défis, cette 6ème édition de Ready For IT se positionne comme un accélérateur de transformation pour nos entreprises.

Ready For IT 2025 : le rendez-vous stratégique des décideurs IT

Du 20 au 22 mai à Monaco, la 6ème édition de Ready For IT réunira les DSI, CTO, RSSI et Directeurs Innovation des ETI françaises, ainsi que les principaux offreurs de solutions en transition et sécurité numériques. Cet événement s'impose comme une plateforme d'échanges incontournable, permettant aux décideurs de structurer leur transformation numérique dans un contexte marqué par de nouvelles exigences réglementaires et économiques.

En 2025, les DSI doivent composer avec de nouvelles contraintes réglementaires (NIS2), une exigence accrue en matière de RSE et des défis RH toujours plus complexes. La transformation digitale ne se limite plus à un levier d'innovation : elle est désormais un impératif de compétitivité internationale et de maîtrise des risques. Dans ce cadre, les entreprises doivent optimiser leurs processus, rationaliser leurs coûts et renforcer leur cybersécurité pour assurer un développement pérenne.

Le programme de Ready For IT 2025 met en lumière ces enjeux à travers des conférences, des tables rondes et des rencontres One to One, offrant aux participants une vision stratégique et des solutions concrètes pour réussir leur transition numérique.

Anticiper l'avenir et structurer la transformation numérique

Face aux évolutions rapides du numérique et aux impératifs de sécurité, nos entreprises doivent adopter une approche proactive et s'appuyer sur des écosystèmes solides pour assurer leur résilience. Ready For IT 2025 sera l'occasion d'échanger sur les meilleures pratiques, d'identifier les leviers d'action et de bâtir des stratégies robustes pour faire de la transformation numérique un moteur de croissance durable.

Ne manquez pas cet événement clé pour anticiper les mutations du numérique et sécuriser la transformation de votre organisation. Rendez-vous à Monaco du 20 au 22 mai !

Scannez ce QRcode pour en savoir plus

L'OBSERVABILITÉ FULL STACK

FOCUS

L'Estonie en **opération séduction** en France

L'ambassade d'Estonie a inauguré, lundi 10 mars, un pôle d'affaires au sein de son ambassade, qui doit faciliter l'implantation des entreprises estoniennes en France et attirer les investissements français en Estonie.

Rendez-vous était pris à l'ambassade d'Estonie, en présence de Lembit Uibo, ambassadeur d'Estonie en France, de Margus Tsahkna, ministre des Affaires étrangères d'Estonie, et de Benjamin Haddad, ministre chargé de l'Europe, pour l'inauguration d'un pôle d'affaires.
Intégré au sein même de l'ambassade, l'objectif est double, a rappelé Lembit Uibo : *« Ce business hub est destiné, tout d'abord, bien sûr, à accueillir les entreprises estoniennes intéressées par le marché français, les accompagner dans leurs recherches de partenaires locaux et dans leur implantation. »* Le petit pays balte espère en outre, par ce biais, attirer les capitaux français sur les côtes de la mer Baltique. *« Mais aussi, inversement, lorsqu'il y a un intérêt du côté français, vous trouverez des moyens de vous diriger vers la bonne entreprise en Estonie. Et, bien sûr, si vous êtes intéressé par un investissement en Estonie »*, a-t-il ajouté.
Concrètement, dans ce hub, les entreprises pourront organiser des rendez-vous d'affaires et des présentations de produits avec leurs partenaires. De son côté, l'ambassade organisera des réunions dédiées aux conseils juridiques, financiers, stratégiques…

L'Estonie, un environnement favorable aux entreprises

Les relations commerciales entre l'Estonie et la France ont connu une forte croissance, notamment dans les échanges de services (+22 %). Une croissance largement portée par le secteur du numérique, un domaine maîtrisé par la petite république balte, qui a rendu disponibles en ligne 100 % de ses services publics, et qui est souvent affublée du sobriquet de *« Silicon Valley de l'Europe »*, notamment en raison d'un environnement favorable, que n'a pas manqué de rappeler l'ambassadeur. *« Notre société numériquement avancée, notre écosystème économique flexible et un environnement favorable aux entreprises ont transformé l'Estonie en un hub mondial de l'innovation. Et nous sommes véritablement aussi une nation de startups. »*
Le pays s'est notamment doté, depuis 2014, d'un programme de e-résidence, qui permet aux personnes n'ayant pas la nationalité estonienne de bénéficier

© Pascal Yuan

Lembit Uibo, ambassadeur d'Estonie en France

d'une identité numérique leur donnant accès à des services de création d'entreprises, bancaires et fiscaux, notamment. Actuellement, l'Estonie compte 120 000 e-résidents qui ont créé 34 000 entreprises. *« Tout est en ligne […] Vous pouvez vraiment vous concentrer davantage sur la création de votre entreprise que sur la bureaucratie […] dans un environnement qui est également le meilleur écosystème fiscal d'Europe […] Nous n'avons pas de taxe sur les sociétés »*. (Le pays applique toutefois un impôt de 20 % sur les dividendes) ont tenu à rappeler le ministre et l'ambassadeur.
Le pays compte actuellement 1 500 startups qui ont généré plus de 2,1 milliards d'euros de chiffre d'affaires en 2021. Et, depuis 2010, la communauté dynamique des startups estoniennes a levé un total de plus de 4,1 milliards d'euros, énonce Lembit Uibo. En plus d'une dizaine de licornes, telles que la plateforme de mobilité Bolt, le pays compte une douzaine d'autres jeunes pousses évaluées à plus de 100 millions d'euros chacune. Le message est clair : *« Alors, s'il vous plaît, faisons des affaires et nous serons plus forts »*, s'est exclamé Margus Tsahkna. Le ton est donné.

© Pascal Yuan

Margus Tsahkna, ministre des Affaires étrangères d'Estonie, et Benjamin Haddad, ministre chargé de l'Europe

TSMC investit 100 Md$ dans les puces aux États-Unis

Le géant taïwanais TSMC a annoncé un nouvel investissement de 100 milliards de dollars dans la fabrication de semi-conducteurs sur le sol américain. Cette nouvelle enveloppe vient s'ajouter aux 65 milliards de dollars déjà prévus pour la production de puces avancées dans des usines situées à Phoenix, en Arizona. TSMC prévoit de construire trois nouvelles usines de fabrication, comprenant deux installations de conditionnement avancé et un centre de R&D. Actuellement, TSMC exploite une autre usine à Camas, dans l'État de Washington, ainsi que des centres de conception à Austin (Texas) et San Jose (Californie).

L'industriel assure que cet investissement générera des centaines de milliards de dollars de valeur dans le secteur des semi-conducteurs, et créera 40 000 emplois dans la construction, ainsi que des dizaines de milliers d'autres, hautement qualifiés, dans la fabrication avancée de puces et la R&D. L'investissement devrait également générer 200 milliards de dollars de retombées économiques indirectes en Arizona et aux États-Unis au cours des dix prochaines années promet le Taïwanais.

« En 2020, grâce à la vision et au soutien du président Trump, nous avons entrepris notre voyage pour établir une fabrication avancée de puces aux États-Unis. Cette vision est aujourd'hui une réalité », a déclaré le président et PDG de TSMC, C.C. Wei.

Avant Donald Trump, TSMC avait déjà pu compter sur le soutien de l'administration américaine, à hauteur de 6,6 milliards de dollars dans le cadre du Chips Act, un plan de 52 milliards de dollars lancé par Joe Biden, visant à créer une chaîne d'approvisionnement aux États-Unis et à réduire la dépendance du pays vis-à-vis des acteurs étrangers. À noter que trois usines TSMC avaient déjà été annoncées en avril 2024. Une première usine en Arizona a démarré la production de puces en 4 nm. La seconde produira des technologies de gravure en 2 nm, ainsi que la technologie 3 nm à partir de 2028. La troisième se concentrera sur la production de composants gravés en 2 nm et moins d'ici 2030. Le Chips Act a également bénéficié à d'autres acteurs du secteur, tels qu'Intel ou encore GlobalFoundries.

Tout comme son prédécesseur démocrate, Joe Biden, Donald Trump cherche à relocaliser la production de puces aux États-Unis afin d'en sécuriser l'approvisionnement. De leur côté, les autorités taïwanaises ont exigé que les processus de fabrication les plus avancés restent à Taïwan. L'île cherche en effet à préserver son « bouclier de silicium ». Sa position de premier fabricant mondial de puces dissuade Pékin d'engager une action militaire visant à rattacher Taïwan à la Chine. En effet, une telle action risquerait d'affaiblir la chaîne d'approvisionnement chinoise et, plus largement, l'économie mondiale. Les États-Unis et d'autres puissances mondiales pourraient ainsi être plus enclins à défendre Taïwan en raison de son importance stratégique.

OpenAI et CoreWeave s'affranchissent un peu plus de Microsoft

À l'approche de son introduction en bourse, la startup d'intelligence artificielle CoreWeave frappe un grand coup en signant un accord de 11,9 milliards de dollars sur cinq ans avec OpenAI. Concrètement, l'entreprise mettra son infrastructure d'IA au service d'OpenAI. De son côté, la maison mère de ChatGPT recevra 350 millions de dollars d'actions CoreWeave.

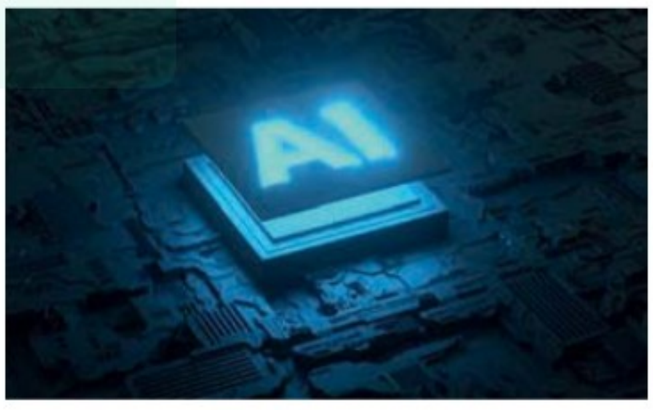
Pour rappel, CoreWeave exploite un service cloud spécialisé dans l'IA, s'appuyant sur un réseau de 32 centres de données dotés d'au moins 250 000 GPU Nvidia. L'entreprise a également ajouté à son infrastructure la dernière architecture de Nvidia, Blackwell.

« CoreWeave est un ajout important au portefeuille d'infrastructures d'OpenAI, complétant nos accords commerciaux avec Microsoft et Oracle, ainsi que notre coentreprise avec SoftBank sur Stargate », a déclaré Sam Altman, PDG d'OpenAI.

Trouver d'autres financeurs

Au-delà de cette déclaration, ce contrat marque une nouvelle étape dans l'émancipation d'OpenAI vis-à-vis de Microsoft, mais aussi celle de CoreWeave, historiquement très dépendante de la firme de Redmond, qui représentait encore 62 % de ses revenus en 2024. L'accord est également perçu comme un moyen pour CoreWeave de rassurer les investisseurs avant son entrée en bourse en diversifiant ses sources de revenus, elle qui vise une valorisation à 32 milliards de dollars.

Pour OpenAI, l'intérêt est ailleurs. Bien que Microsoft ne soit plus son fournisseur exclusif, c'est principalement



son infrastructure cloud (y compris via CoreWeave) qui lui a permis d'entraîner ses modèles et de maintenir ses services. Mais Microsoft ne suffit plus, et la startup est en quête permanente de nouvelles ressources informatiques indispensables à ses besoins croissants. En janvier 2025, OpenAI annonçait une collaboration avec Oracle et SoftBank dans le cadre du projet Stargate. Cet investissement de 500 milliards de dollars aux États-Unis doit construire de nouvelles infrastructures dédiées à l'IA, qui seront exploitées par OpenAI.

RACHATS

ServiceNow s'offre la startup d'IA **Moveworks** pour 2,9 Md$

Expert de l'IA agentique et de l'automatisation des flux de travail complexes, ServiceNow a annoncé, ce lundi, le rachat de son concurrent Moveworks pour 2,85 milliards de dollars. Un record pour l'éditeur de logiciels. Moveworks développe un assistant d'IA agentique qui connecte les systèmes d'entreprise pour optimiser les flux de travail des employés de clients tels que Broadcom, Palo Alto Networks et Siemens, pour ne citer qu'eux. Son outil peut être intégré à ServiceNow, mais aussi à Slack (Salesforce) et SharePoint (Microsoft), et compte actuellement 5 millions d'utilisateurs.

Avec cette acquisition, ServiceNow veut combiner ses capacités d'IA agentique et d'automatisation avec l'assistant d'IA et la technologie de recherche d'entreprise de Moveworks pour fournir aux employés un assistant d'IA universel, capable de fournir des réponses rapides et d'automatiser les tâches répétitives. *« Aujourd'hui, la majorité des clients de Moveworks utilisent déjà ServiceNow comme un système clé pour accéder à l'IA d'entreprise, aux données et aux flux de travail, ce qui garantira une intégration fluide entre les deux entreprises »*, précise un communiqué. Les deux entreprises comptent d'ailleurs 250 clients communs.

SolarWinds met la main sur **Squadcast**

SolarWinds étend les capacités de sa plateforme. Le spécialiste de l'observabilité et de l'ITSM vient de mettre la main sur la société Squadcast qui édite une solution de réponse aux incidents. L'objectif de cette acquisition, dont le montant n'a pas été communiqué, est de permettre à SolarWinds d'accélérer les temps moyens de résolution des incidents (MTTR) pour les utilisateurs de sa plateforme.

Avec cette acquisition, SolarWinds vient ajouter un nouvel atout à sa solution de gestion des environnements IT. Selon Squadcast, les utilisateurs de sa technologie constatent une réduction de 68 % du MTTR moyen, une économie de quelque 1 000 heures de travail et de 500 000 dollars grâce à sa technologie de réponse aux incidents.

Le spécialiste de la formation **Cyber Guru** reprend **Mantra**

L'italien Cyber Guru va absorber les équipes et clients de Mantra, une startup française qui développe des solutions de sensibilisation des collaborateurs à la cybersécurité.

Fondée en 2017, Cyber Guru développe une plateforme de formation et de sensibilisation à la sécurité pour les entreprises, alimentée par un modèle d'IA propriétaire basé sur l'apprentissage automatique. Elle compte 700 clients et un million d'utilisateurs uniques répartis dans plus de 90 pays.

Avec cette acquisition, elle absorbera les équipes et les outils de Mantra, et, au final, touchera près de 1000 organisations et 1,5 million d'utilisateurs actifs. Cyber Guru souhaite combiner ses moteurs d'apprentissage automatique à ceux de Mantra, pour atteindre plus de 340 milliards de points de données qui viendront nourrir les modèles d'apprentissage automatique, permettant ainsi de générer des formations plus personnalisées. L'entreprise a également indiqué vouloir accélérer sa croissance au cours des 12 prochains mois, en s'appuyant sur des initiatives stratégiques et des solutions innovantes. Elle a précisé vouloir renforcer ses effectifs en France et en Italie, et se concentrer sur le marché français en augmentant ses investissements en R&D.

SoftBank rachète **Ampere Computing** pour 6,5 milliards de dollars

La multinationale japonaise SoftBank multiplie les opérations dans le domaine des infrastructures d'IA aux États-Unis. Après avoir annoncé en début d'année sa participation dans le projet pharaonique de Donald Trump, Stargate, avec OpenAI, Oracle et le fonds d'investissement MGX, elle a révélé, jeudi 20 mars, avoir mis la main sur l'entreprise de conception de microprocesseurs haute performance pour centres de données, Ampere Computing. Montant de l'opération : 6,5 milliards de dollars. Un gros coup pour le Japonais, dans un contexte de forte demande d'infrastructures dédiées au calcul d'IA. Ampere va être amenée à soutenir l'écosystème de SoftBank et les entreprises du groupe, notamment dans le développement et la fabrication de puces basées sur Arm, dont elle est actionnaire majoritaire. SoftBank espère finaliser l'opération dans la seconde moitié de 2025, sous réserve des approbations réglementaires, notamment des autorités antitrust et du Comité sur les investissements étrangers aux États-Unis. La transaction a d'ores et déjà été approuvée par le conseil d'administration de SoftBank. Ampere conservera son nom et opérera comme filiale de SoftBank.

Anthropic lève 3,5 Md$

Le concurrent d'OpenAI, Anthropic, derrière le LLM Claude, a annoncé avoir bouclé une Série E de 3,5 milliards de dollars. Cette opération a été menée par Lightspeed Venture Partners, avec la participation de Bessemer Venture Partners, Cisco Investments, D1 Capital Partners, Fidelity Management & Research Company, General Catalyst, Jane Street, Menlo Ventures et Salesforce Ventures, ainsi que d'autres investisseurs nouveaux et existants. Elle porte la valorisation de l'entreprise à 61,5 milliards de dollars. Avec ces fonds, le concurrent de ChatGPT entend accélérer le développement de ses systèmes d'IA, augmenter sa capacité de calcul et accélérer son expansion internationale.

Côté technologie, la course aux modèles fait rage. Claude 3.7, Grok 3, OpenAIo1, GPT-4.5, Mistral Saba, Le Chat de Mistral… Chaque acteur du secteur enchaîne les lancements ces derniers mois, à grands renforts de benchmarks et de nouvelles capacités annoncées, notamment en raisonnement. Tous vantent les capacités supérieures ou au moins égales de leurs modèles face à ceux des concurrents.

Dexterity AI engrange 95 millions de dollars pour ses robots industriels

La startup américaine de robotique industrielle Dexterity AI a levé 95 millions de dollars. Une opération qui porte sa valorisation à 1,65 milliard de dollars. Dexterity AI met au point ce qu'elle appelle une « IA physique ». Elle anime ses robots dotés de « dextérité humaine », dédiés aux applications industrielles et plus particulièrement à la logistique, telles que le chargement et le déchargement de camions et de conteneurs, le tri, la palettisation et la dépalettisation. Ses technologies sont déjà utilisées par des entreprises telles que FedEx et UPS.

Son robot en développement, baptisé Superhumanoïde Mech, est présenté comme le premier du genre par l'entreprise. Il est capable de soulever des charges allant jusqu'à 60 kg, dispose d'une envergure de bras de 5 mètres et peut résister à des températures extrêmes. Son IA physique lui confère une vision et un sens du toucher.

Quantum Machines empoche 170 M$

Quantum Machines a annoncé une levée de 170 millions de dollars en Série C. L'opération a été dirigée par PSG Equity, avec la participation d'Intel Capital, de Red Dot Capital Partners et d'investisseurs existants.

Quantum Machines développe notamment une technologie de contrôle hybride permettant de gérer efficacement des calculs complexes sur différents types d'ordinateurs quantiques. Elle collabore avec Nvidia via la plateforme de calcul et de contrôle quantique Nvidia DGX Quantum, utilisée pour la correction des erreurs.

L'entreprise affirme que 50 % des acteurs développant des ordinateurs quantiques utilisent ses technologies pour concevoir et faire évoluer leurs systèmes.

500 millions de dollars pour NinjaOne

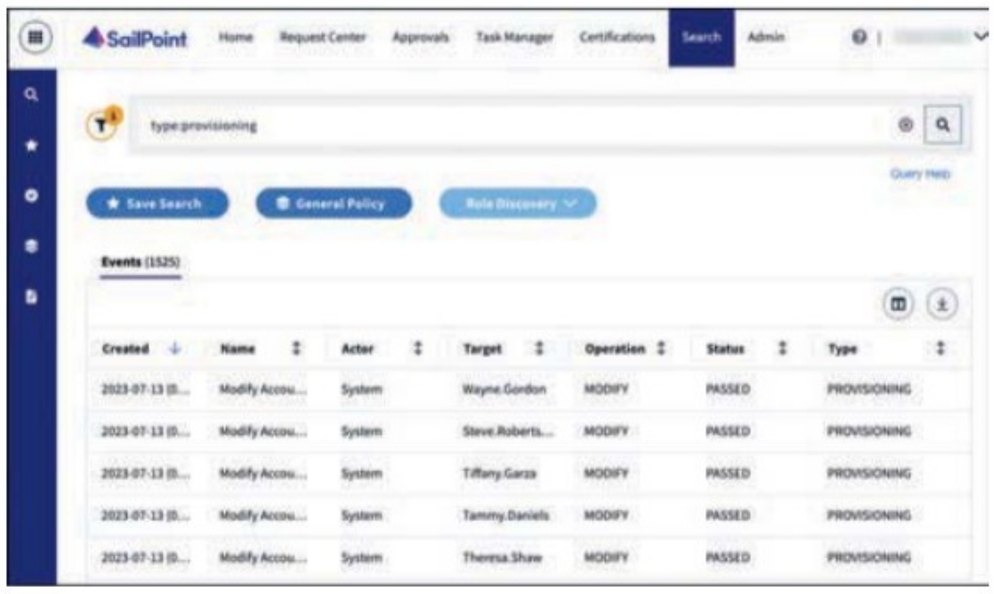
La plateforme de gestion automatisée des terminaux a annoncé une extension de sa série C de 500 millions de dollars, ce qui porte sa valorisation à 5 milliards de dollars. *« Le financement en capital-risque a été réalisé en plusieurs tranches distinctes, menées par Iconiq Growth, CapitalG, le fonds d'investissement indépendant d'Alphabet, et d'autres investisseurs privés de premier plan »*, a précisé l'entreprise dans un communiqué.

NinjaOne développe des solutions de gestion des terminaux, des appareils mobiles (MDM), d'accès à distance, de sauvegarde et de sécurité. Elle compte parmi ses 24 000 clients, Nvidia, Nissan, Cintas, Vimeo et HelloFresh.

Avec ces fonds, l'entreprise entend dynamiser sa R&D dans la gestion autonome des terminaux, l'application autonome de correctifs, la correction des vulnérabilités et l'amélioration de son expérience client. Une partie de cette somme sera également fléchée vers le financement de l'acquisition en cours de Dropsuite, une entreprise spécialisée dans la sauvegarde SaaS et la protection des données, qu'elle souhaite racheter pour 262 millions de dollars.

Sailpoint étend son offre vers les MSP

Le spécialiste de la cybersécurité étend son programme aux partenaires MSP qui pourront offrir leurs services à un plus large éventail d'entreprises.
L'expansion du programme MSP de SailPoint, initialement lancé en 2024, permet à davantage d'entreprises d'intégrer la sécurité des identités en tant que fonction clé en se concentrant sur les use cases de premier niveau. Parallèlement, cette expansion crée des opportunités pour développer et faire évoluer le programme par la suite. Grâce aux partenaires MSP actuels (et d'autres à venir), cette approche permet aux entreprises de faire évoluer leurs capacités d'identités au rythme de leur croissance.
Cette initiative de mise sur le marché exclusive aux partenaires permet à davantage de partenaires MSP d'utiliser les solutions SailPoint, afin de créer des solutions sur mesure pour les petites entreprises. Déjà fournisseur auprès des grandes entreprises, Sailpoint pourra, grâce à cet investissement



élargi, permettre à un plus grand nombre d'entreprises - plus petites, mais aussi plus nombreuses - d'exploiter le potentiel de ses solutions de sécurité des identités.

Cisco et Ineso unis dans le projet Valomilk

Le fournisseur de solutions réseau et de sécurité se rapproche de l'intégrateur de solution d'Internet des Objets dans le cadre du projet Valomilk, une initiative visant à digitaliser l'élevage laitier en France.
Lancé en janvier 2024, ce partenariat stratégique, en association avec Rhône Conseil Élevage, a déjà permis d'apporter des solutions innovantes et mesurables pour améliorer la productivité, le bien-être animal et la durabilité des exploitations agricoles. Le projet a pour but de démontrer comment les technologies de pointe peuvent être des leviers de transformation pour l'agriculture française. Au cœur du dispositif, l'intégration de capteurs IoT, combinée à l'intelligence artificielle et à la blockchain, permet d'analyser en temps réel les données de santé et de comportement des animaux. Cette approche améliore la traçabilité des données, optimise les pratiques agricoles et renforce la prise de décision des éleveurs.

Des premiers résultats probants

Depuis son lancement, Valomilk a déjà permis d'obtenir des avancées significatives sur le terrain, avec une baisse des températures de 10 C° dans les bâtiments rénovés, sans climatisation, améliorant ainsi le confort animal et réduisant les coûts énergétiques. Il a aussi été constaté une augmentation des revenus des exploitations entre 10 et 15 % grâce à l'optimisation des processus et une meilleure gestion des ressources. Par la surveillance en temps réel de l'état de santé des animaux, les maladies infectieuses ont connu une décrue de 10 %.
En 2025, le projet prévoit d'étendre son dispositif à 600 fermes, permettant ainsi de surveiller 39 000 vaches et d'analyser 200 millions de points de données. Cette expansion vise à renforcer les modèles prédictifs et améliorer la résilience des exploitations face aux aléas climatiques. En parallèle, Cisco et Ineso poursuivent leurs efforts pour intégrer davantage d'éleveurs dans cette dynamique, en proposant des solutions adaptées à tous les types d'exploitations, tout en garantissant une cybersécurité robuste et un accès fiable aux infrastructures numériques.

Check Point et Cardano partenaires autour de la Blockchain

Le fournisseur de solutions de sécurité a désormais un partenariat stratégique avec Cardano, une plateforme blockchain reconnue pour son engagement en faveur de la durabilité, de l'évolutivité et de la transparence.
Les deux acteurs ambitionnent de renforcer durablement la confiance des institutions et des grandes entreprises sur la technologie de chaîne de blocs. Ensemble, ils souhaitent proposer la toute première solution complète de sécurité blockchain en temps réel. Ce partenariat stratégique a pour objectif de combiner l'expertise de pointe de Check Point en cybersécurité avec la robustesse éprouvée de l'infrastructure blockchain développée par Cardano. Check Point mettra à disposition du réseau Cardano ses technologies avancées de surveillance et de renseignement sur les cybermenaces. Cela inclut notamment des outils de détection et de prévention en temps réel, capables d'identifier et de neutraliser un large éventail d'attaques, allant des vulnérabilités affectant les contrats intelligents jusqu'aux tentatives de phishing.
Pour sa part, Cardano apporte à ce partenariat une expertise technique éprouvée et une infrastructure solide. Cette collaboration permettra d'intégrer les capacités de cybersécurité de Check Point, et assurera une surveillance complète et continue de toutes les opérations effectuées sur sa blockchain.

Juniper et **IBM** main dans la main pour la sécurité des réseaux

Les deux sociétés annoncent le développement de leur partenariat en matière de ventes, de marketing et d'intégration de produits. Les entreprises projettent de réunir les forces de Mist AI de Juniper et d'IBM watsonx pour la gestion des réseaux IT, améliorer l'expérience des utilisateurs et réduire les coûts d'exploitation.
Cette initiative se concentrera sur le déploiement de Mist AI et d'IBM watsonx. Ainsi, IBM et Juniper travailleront sur deux projets internes d'IBM dont IBM Guest Service, un projet qui intègre Mist AI avec IBM watsonx Orchestrate, afin d'automatiser l'assistance au réseau IT, de réduire le nombre de tickets d'assistance Wi-Fi et de résoudre les incidents liés au réseau Wi-Fi des clients. Le second projet est un outil interne en cours de développement par le département IT d'IBM qui utilise IBM watsonx Orchestrate pour s'intégrer à des outils comme Mist AI de Juniper, et résoudre les problèmes d'infrastructure réseau en diagnostiquant les problèmes des utilisateurs finaux. Appelé AskNetwork, il devrait ainsi simplifier les données techniques complexes pour les transformer en informations exploitables, afin d'aider les entreprises à optimiser les performances de leur réseau.
Juniper et IBM envisagent également d'explorer d'autres fonctionnalités, comme les renseignements et les diagnostics basés sur la localisation, capables d'apporter une valeur ajoutée à la gestion des installations et à l'expérience utilisateur des patients et des clients dans des secteurs tels que la santé et le retail.

Orange se renforce dans la connectivité satellite

Après son partenariat avec Eutelsat sur le continent africain, Orange renforce encore sa solution de connectivité par satellite avec un nouvel accord avec Telesat.
Dans le cadre de cet accord, une station d'atterrissage Telesat Lightspeed sera hébergée au téléport d'Orange à Bercenay-en-Othe (10) et bénéficiera d'une connexion terrestre avec leur Point de Présence (PoP) situé à Paris, via une Ligne Privée Internationale (IPL) fournie par Orange Wholesale. De plus, Orange a signé un engagement de capacité lui permettant d'intégrer le service Telesat Lightspeed en Orbite Terrestre Basse (OTB) dans son portefeuille mondial de services pour les entreprises et les opérateurs de télécommunications. Ce dernier réseau va renforcer la résilience pour le « mobile backhauling », les situations d'urgence et la connectivité à distance.
Les opérateurs de télécommunications peuvent configurer et contrôler les services en temps réel, redirigeant si nécessaire la capacité sans intervention de l'opérateur satellite. Ce partenariat renforcera la capacité d'Orange à répondre aux besoins des clients dans les régions mal desservies, ajoutant une nouvelle option de connectivité.

Inetum et **Workday** signent un partenariat

Inetum souhaite se positionner comme un intégrateur et revendeur majeur pour Workday en France.
Ce partenariat s'inscrit dans la stratégie d'expansion de Workday en Europe. En rejoignant le programme Workday Extend, Inetum va développer des solutions innovantes pour les RH et la finance, disponibles sur la marketplace Workday. L'accord ouvre également à Workday le marché des acteurs publics et semi-publics, tant auprès des institutions et ministères que des collectivités ou du monde médical. Workday bénéficiera, par ailleurs, de la proximité qu'Inetum entretient avec ses clients à l'échelle locale pour développer sa visibilité en régions sur le territoire français.

AGENDA

Big Date and AI World
9 - 10 avril 2025
Recinto Montjuic, Barcelone, Espagne

Google Cloud Next
9 - 11 avril 2025
Mandalay Bay Convention Center, Las Vegas USA

Veeam ON
21 - 23 avril 2025
Marriott Marquis San Diego Marina, San Diego USA

RSA Conference
28 avril - 1er mai 2025
Moscone Center, San Francisco USA

IBM Think
5 - 8 mai 2025
Hynes Convention Center, Boston USA

Nutanix Next
7 - 9 mai 2025
Walter E. Washington Convention Center, Washington USA

IoT World Congress
13 - 15 mai 2025
Gran Via Venue | HALL 8, Barcelone Espagne

Une plateforme unifiée de sécurité et d'observabilité pour une résilience inégalée.

De nombreuses organisations parmi les plus grandes et complexes au monde s'appuient sur Splunk pour contribuer à assurer la sécurité et la fiabilité de leurs systèmes numériques. Découvrez notre plateforme unifiée de sécurité et d'observabilité sur splunk.com/fr_fr

splunk>
a CISCO company

Observabilité

La performance par les données

Le paysage informatique des entreprises n'arrête pas de se complexifier. Il devient difficile pour les entreprises de suivre le rythme sans une stratégie mature pour l'IA, l'analyse et l'automatisation nécessaire. Pour conserver la performance de leur pile informatique allant de l'infrastructure aux applications, l'observabilité a un rôle important à jouer pour gérer les incidents avant qu'ils n'impactent réellement les utilisateurs. L'observabilité, à la différence des outils classiques de monitoring, ne vous dit pas où et quel élément est défectueux, mais vous donne le pourquoi de l'incident et vous indique quels utilisateurs et quels services critiques de votre entreprise sont touchés. Si le concept peut sembler ancien, il connaît un large renouveau principalement pour les environnements en cloud et les applications nativement cloud.

Son principal bénéfice est de s'appuyer sur les données, même si leur volume et leur analyse peut sembler une difficulté parfois difficile à surmonter. Ce dossier essaie de faire le point sur l'utilisation et la maturité des entreprises sur l'observabilité, ainsi que les outils à leur disposition pour la mettre en œuvre.

Dossier réalisé par Bertrand Garé

Une **maturité** encore faible

En général, les entreprises et les vendeurs ont du mal à s'accorder sur le concept d'observabilité suivant qu'ils viennent du réseau, du suivi d'applications, de la sécurité… De la même manière, les clients utilisent le mot observabilité avec de nombreux sens différents suivant leur maturité et leurs besoins.

Stéphane Estevez, EMEA Observability Market Advisor chez Splunk, n'y va pas par quatre chemin : « *Aujourd'hui, tout le monde parle d'observabilité, tout le monde utilise les mêmes mots clés. Chez mes confrères ou même chez Splunk, on est dans le même cas, tout le monde dit : C'est du 'end-to-end' de visibilité, c'est du real-time… Mais personne ne dit comment.*

Au final, c'est impossible pour un DSI de prendre une décision sans être obligé de gratter vraiment la surface. Donc, un des problèmes, c'est la terminologie. Quand les clients eux-mêmes parlent d'observabilité, souvent, il s'agit de télémétrie, de monitoring. Ce n'est pas vraiment de l'observabilité. Que ce soit chez les vendeurs, chez les intégrateurs, chez les clients, personne n'est d'accord sur ce que c'est. »

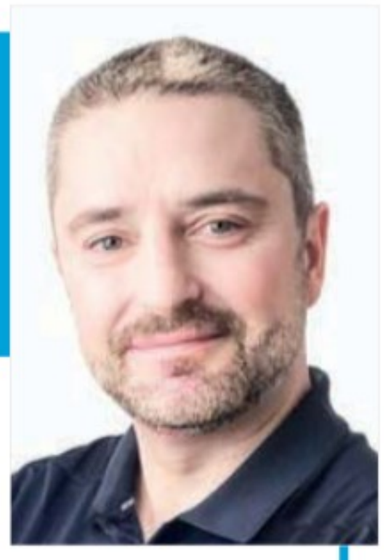
Alexandre Signoret, en charge de l'offre d'observabilité et AIOPs chez IBM, est du même avis. « *Je sens exactement la même chose. Je pense que le marché est confus, et il est confus parce que je pense que les clients sont encore à la recherche de la direction que ça va prendre. Les éditeurs aussi ont leur part de responsabilité par leur message marketing et l'approche qu'ils ont aujourd'hui, on a l'impression que l'observabilité est un outil pour répondre à tout en essayant de faire tout rentrer dedans.* »

Thomas di Luccio, Product Manager chez Platform.sh mais venant de BlackFire, un outil d'observabilité racheté par Platform.sh, perçoit deux types de population face à l'observabilité : « *Il y a les pompiers, ceux qui n'ont pas investi dans l'observabilité. Ils ont vu la performance comme un* « nice to have », *un truc en plus dont on peut se passer, parce que l'énergie doit se mettre à créer des fonctions. Et forcément, il y a eu un lancement de produits, il y a eu une offre commerciale, Black Friday, des offres saisonnières, et tout s'est effondré. On a perdu énormément d'argent avec une vente qui s'est loupée, et ils viennent parce qu'il y a une crise. Il y a une autre catégorie de gens, malheureusement moins nombreux, qui ont une vision plus proactive de la relation à la performance. Donc ils viennent à l'observabilité par la question de la performance, du contrôle. Ils disent : « OK, la performance est une fonction intégrante du produit qu'on crée, donc on a besoin d'être en contrôle. » Et pour être en contrôle, on a besoin de comprendre qu'est-ce qui peut bien se passer dans nos applications, qu'est-ce qui peut bien se passer dans nos serveurs qui font que, de temps en temps, ça frictionne, on n'arrive pas à comprendre pourquoi, et on a besoin de cette capacité à observer les systèmes. En tout cas, moi, je le définis comme ça. Je sais qu'il y a un flou là-dessus. Moi, je le définis comme une exception assez large. C'est une capacité. On offre la capacité à voir dans les systèmes.* »

Une utilisation émergente

Si le sujet fait beaucoup parler et écrire, les entreprises entament juste le chemin vers l'observabilité. Selon une étude réalisée pour le compte d'OpsRamp, une société dans le giron d'HPE, 30 % des répondants indiquent explorer les cas d'usages adéquats pour l'observabilité. Seulement 24 % ont mis en œuvre une suite d'outils d'observabilité complète dans plus de 90 % de leur organisation, dont 19 % que dans certaines entités de l'entreprise. Les outils en place servent plus particulièrement à observer les applications dans le cloud ou nativement cloud (61 %). La moitié les utilise dans des environnements hybrides ou pour des questions de sécurité. Un

Stéphane Estevez,
EMEA Observability Market Advisor chez Splunk



« *Aujourd'hui, tout le monde parle d'observabilité, tout le monde utilise les mêmes mots clés. Au final, c'est impossible pour un DSI de prendre une décision sans être obligé de gratter vraiment la surface. Donc, un des problèmes, c'est la terminologie* »

peu moins (47%) l'utilisent pour suivre le réseau. 34% indiquent cependant suivre l'ensemble du système informatique.

Les coûts conditionnent l'usage

Selon la même étude, les coûts de licences mènent l'utilisation. De ce fait, à 49%, les entreprises choisissent une formule à prix fixe dans le cadre d'accord entreprise. Seulement 12% prennent la formule *« pay as you go »*. Du fait du volume de données et autres éléments corollaires, les entreprises font le choix de prévoir les coûts quand elles mettent en place l'observabilité. De plus, les outils sont considérés comme chers et la question des coûts et du retour sur investissement reste centrale dans les projets (53%). Le premier point d'attention reste cependant les données. Le volume des données à gérer et à stocker est trop important pour un usage ou une analyse effective (57%). A la même proportion que les coûts se placent la précision des données et les problèmes de faux positifs. La courbe d'apprentissage des outils avant de pouvoir les utiliser efficacement est aussi un point cité. Viennent ensuite la peur autour de la suppression de certains postes et la longueur des cycles d'implantation.

Un autre point problématique est l'écart entre les bénéfices attendus et la réalité. Le premier bénéfice attendu reste encore l'amélioration de la performance des applications et de l'expérience des utilisateurs. Cela peut s'expliquer par le fait que la plupart des outils du marché proviennent d'acteurs historiques du monitoring de la performance des applications. Ce point est largement devant l'autre bénéfice attendu, l'amélioration de l'efficacité de l'automatisation dans l'ensemble de l'organisation. A égalité suivent une attente sur des déploiements plus rapides des applications avec moins de problèmes et d'arrêt de production et la détection de problèmes complexes. Certains s'attendent même à une détection proactive des problèmes. Plus généralement, ils attendent une meilleure efficacité opérationnelle. Pour beaucoup les objectifs sont atteints. Ainsi, 59% des personnes interrogées dans cette étude indiquent avoir la capacité des problèmes de performance qu'ils ne connaissaient pas et de répondre aux problèmes avant que les utilisateurs soient impactés. Viennent ensuite la possibilité de contrer des attaques, l'amélioration de la performance des applications, le retrait d'applications héritées et la réduction des dépenses du service IT avec une modernisation des applications ou de l'architecture.

UNE **COMPLEXITÉ** TOUJOURS PLUS GRANDE

Une autre étude réalisée pour le compte de Dynatrace peint le contexte et les problématiques que connaissent les entreprises dans des environnements multiclouds. Nous donnons les principaux résultats de l'étude.

- **En moyenne, un environnement multicloud regroupe 12 plateformes et services différents.**
- **92% des organisations affirment que la complexité de leur pile technologique a augmenté au cours des 12 derniers mois et 53% pensent que cette tendance va se poursuivre.**
- **92% des responsables technologiques déclarent qu'en raison de la complexité du multicloud, il est plus difficile de proposer des expériences client exceptionnelles.**
- **88% des responsables technologiques constatent que les piles technologiques cloud-native génèrent des quantités astronomiques de données, bien supérieures à ce que les humains sont capables de gérer.**
- **En moyenne, les organisations utilisent 10 outils d'observabilité ou de monitoring pour gérer les applications, l'infrastructure et l'expérience utilisateur.**
- **81% des responsables technologiques jugent que le temps passé par leurs équipes sur la gestion des outils de monitoring et la préparation des données à analyser bride l'innovation.**
- **86% des responsables technologiques utilisent ou prévoient d'adopter au cours des 12 prochains mois une plateforme unifiée pour les données d'observabilité et de sécurité.**
- **72% des organisations ont déjà adopté l'AIOps pour réduire la complexité de la gestion de leur environnement multicloud et 24% prévoient de le faire au cours des 12 prochains mois.**

Rénover le monitoring

Interrogées sur les outils qu'elles remplaceraient en premier après la mise en place de l'observabilité, les entreprises répondent quasiment aux deux tiers le monitoring du réseau qui arrive largement devant les outils de gestion de la performance des applications, du suivi de l'infrastructure, du cloud ou de l'expérience utilisateur. En fait, les entreprises utilisent l'observabilité, non pas pour remplacer les outils de monitoring en place, mais le complète pour le rendre meilleur. Les entreprises réalisent cette opération par le biais d'intégration avec les outils de monitoring IT, les logiciels d'automatisation des processus, la gestion des événements ou des incidents et l'AIOps. Ceux qui profitent le plus de cette mise en œuvre sont les équipes d'analyse des données et le management devant les équipes opérant le cloud. Suivent les équipes de sécurité et les équipes de développeurs, que ce soient les DevOps ou les développeurs classiques d'applications.

Des données plus critiques que d'autres

Les entreprises ayant répondu à l'étude d'Opsramp placent en tête les métriques comme étant les données les plus critiques dans leur travail (charge CPU, usage mémoire, taux d'erreur systèmes...). De nouvelles s'ajoutent désormais comme le calcul des coûts du cloud, la consommation énergétique. En fait, tout ce qui peut se mesurer en chiffres est appelé à devenir une métrique. Les logs suivent de près devant les événements et les traces. La plus faible présence des traces comme données critiques résulte de la moindre présence d'applications en micro-services ou serverless. En conséquence, les applications créent moins de traces. Si cela semble être le futur, l'étude indique que les entreprises adoptent ce nouveau type d'application lentement et ont actuellement moins besoin des traces que d'autres indicateurs comme les métriques ou les logs. □

B.G

De **nombreux outils** pour des tâches différentes

Selon leur cœur de métier d'origine, les outils d'observabilité servent des objectifs différents. Suivant les besoins et objectifs attendus, il est nécessaire de faire une étude précise sur les outils qui ne font forcément pas tout.

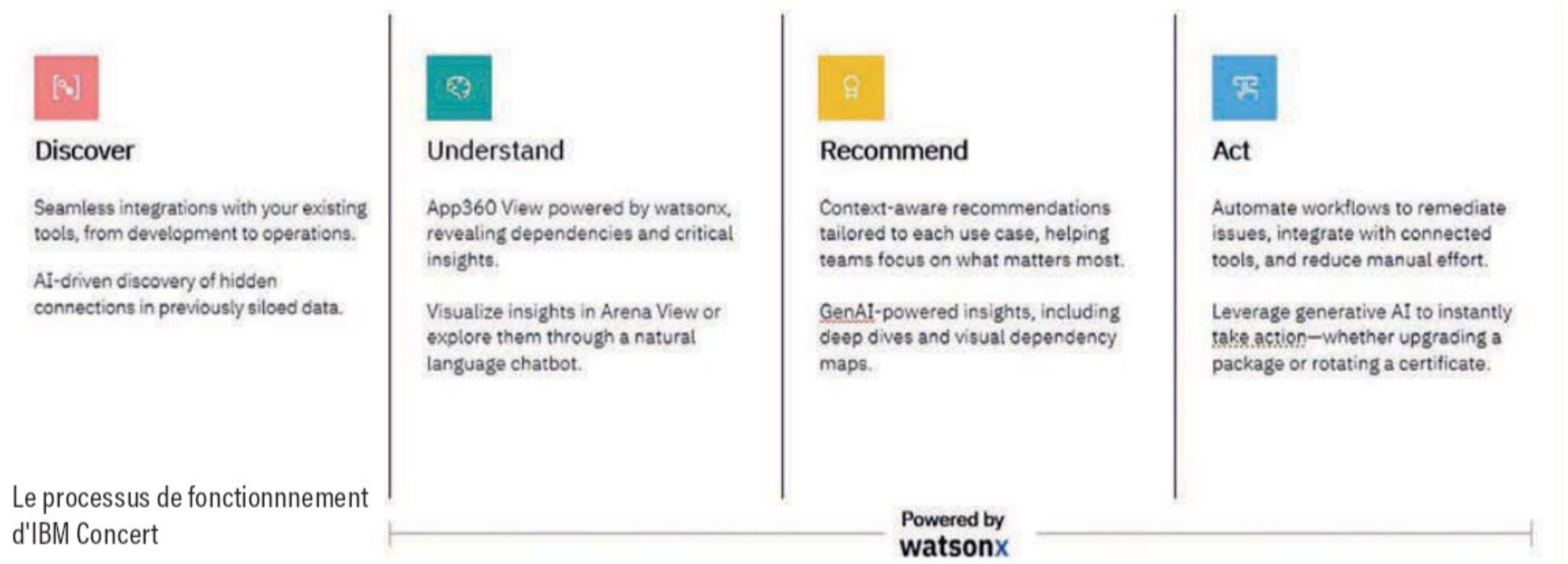
Les outils présents sur le marché ont des origines et des fonctions souvent très différentes. Nous présentons ici quelques exemples de ce qu'il est possible de trouver sur le marché sous le vocable d'Observabilité. Yann Samama, Senior Sales Engineer chez Gigamon, indique : *« Notre approche chez Gigamon, c'est vraiment l'observabilité, c'est ce qu'on voit sur le réseau. Nous estimons que l'intérêt de voir ce qui se passe sur le réseau, c'est que ça donne la réalité des choses, et ce n'est pas juste une estimation de ce qu'il peut être »*. Il continue : *« Donc, si on part du principe que le réseau est un peu comme une espèce d'organisme semi-vivant, qui évolue avec sa propre logique intrinsèque, on est obligé de l'observer pour comprendre ce réseau, pour détecter des menaces éventuelles, pour valider son bon fonctionnement et sa pérennité. Le réseau n'est pas fait pour s'auto-observer, pour s'auto-diagnostiquer. Donc, on a besoin d'aller chercher les informations qui transitent, pour avoir la véracité de ce qui se passe en temps réel, pour prendre les décisions qui s'imposent. Le rôle de Gigamon là-dedans, c'est capturer les paquets, les agréger, éventuellement les filtrer ou les modifier pour apporter un peu plus de valeur dans la chaîne, et enfin les envoyer à des outils. Et ces outils, ça peut être des outils d'observabilité classiques, NPM, APM, comme des outils d'observabilité qui sont orientés sécurité. Et l'idée pour nous, c'est d'être agnostiques. C'est-à-dire qu'on fournit la matière première, et vous en faites votre substantifique moelle, et vous la développez de la manière qui vous semble la plus propice à vos utilisateurs. »*

Easyvista se veut plus modeste en jouant la carte de l'intégration avec d'autres outils, et conserve une forte empreinte hyper ou supervision en particulier sur l'infrastructure.

LES **DÉFIS ET CHALLENGES** DE L'OBSERVABILITÉ

Dans une tribune récente, Quentin de Sainte Marie, Lead Solutions Consultant, IT Operations Cloud, OpenText, mettait en exergue les différents points de difficulté dans la mise en œuvre de l'observabilité.

« Bien que l'observabilité offre de nombreux avantages, elle s'accompagne aussi de défis à surmonter. La complexité croissante des systèmes modernes, qui combinent à la fois microservices, conteneurs et cloud, rend parfois difficile l'identification rapide des causes des incidents. De plus, le volume exponentiel de données générées par les systèmes d'information nécessite des solutions capables de les traiter et de les stocker efficacement. Un autre défi réside dans la disponibilité des compétences nécessaires. La mise en place et l'exploitation d'une plateforme d'observabilité demandent une expertise particulière, d'où l'importance de solutions simples à implémenter et à utiliser. L'intégration avec les outils existants peut aussi être un obstacle, mais des connecteurs prêts à l'emploi peuvent considérablement accélérer ce processus. Enfin, l'adoption de l'observabilité exige une évolution de la culture d'entreprise. Pour maximiser son potentiel, il est donc indispensable de réduire voire supprimer les silos organisationnels, de repenser la composition des équipes techniques et d'encourager une collaboration étroite autour des données ».



Le processus de fonctionnnement d'IBM Concert

Chez IBM, c'est un ensemble de logiciels, souvent issus d'acquisition comme Turbonomics, Apptio qui apporte la complétude, la vision et une observabilité à plusieurs facettes. IBM a de plus développé une console d'agrégation des données de différentes applications sources et de corrélation des données, IBM Concert. Le logiciel a des fonctions de découvertes, de compréhension en s'appuyant sur Watsonx. Il peut proposer des recommandations à l'issue d'un prompt dans un outil d'intelligence artificielle générative, et peut prendre des actions de manière autonome ou remédier à un processus défectueux.

Yves Le Berre chez ALE indique : *« On va agréger aussi les logs divers et variés, les logs d'accès aux machines, les logs des applicatifs et on va agréger aussi tout ce qui est traces d'interaction qui vont nous permettre ce qu'on va appeler aussi des APM, et d'avoir une vision sur la performance de la softisation. Un exemple : on fait des services équivalents à ceux qu'on utilise aujourd'hui, Google Meet, Rainbow, c'est un service de collaboration, de communication. Typiquement, on va vérifier entre la requête de monter par exemple une vidéoconférence, on va aller vérifier le KPI, qui est le temps entre le moment où l'utilisateur va déclencher la conférence et où tous les services vont être mis ensemble pour délivrer le service. On va aller monitorer le temps de réponse, ce qui va donner une indication sur la moyenne, sur le temps, sur ce qu'on est capable de faire et si on se rend compte qu'il y a un délai supplémentaire sur la mise en service, par exemple des serveurs de conférences, on va se poser la question est-ce que c'est un phénomène transitoire, une problématique réseau, une problématique data center, est-ce qu'on manque de serveurs dans la zone, est-ce qu'il ne faudrait pas en rajouter d'autres ou est-ce qu'on a une perte de performance sur un des composants ? »*. Il ajoute : *« sur Rainbow, on a environ un peu plus de 700 alarmes qui sont mises en place, qui sont cascadées sur la chaîne des métriques qu'on va concentrer, qu'on va agréger avec Prometheus. On va associer Prometheus avec l'alert manager et on va effectivement povoir définir des seuils ou des scénarii, donc un enchaînement d'évolutions de KPI qui vont déclencher les alertes et enclencher nos équipes d'opération si on dépasse le gabarit »*.

Thomas di Luccio, Developer Relations Engineer chez Platform.sh, fait découvrir un autre prisme de l'observabilité. *« On fait de l'observabilité côté applicatif, donc concentré essentiellement sur les web apps, PHP Python, c'est notre cœur de métier. On commence à gérer un peu plus de runtime, mais ça reste essentiellement des applications découplées PHP Python »*. Il continue : *« nous réalisons du profiling déterministe, du continuous profiling, donc du profiling probabiliste. Cela veut dire que vous pouvez anticiper comment devrait se comporter une application »*. Il ajoute : *« Blackfire va envoyer des requêtes dans l'environnement que vous avez désigné, faire l'ensemble de ses actions, mesurer la performance de ses actions, et évaluer ça par rapport à ce que vous avez défini. Parce que vous avez défini qu'il fallait que ça se passe en tant de temps, que ça consomme autant de requêtes SQL, que ça fasse tant de CPU, tout ce que vous voulez. Et ça, vous allez pouvoir le rejouer tout le temps.*

Une vue de l'outil de Splunk

On est dans la proactivité quand les développeurs ou les équipes de développement vont pouvoir tester ça, et d'interdire de fusionner leur travail, de le mettre en production, si cela dégrade la performance. On offre la possibilité d'être dans le contrôle ».

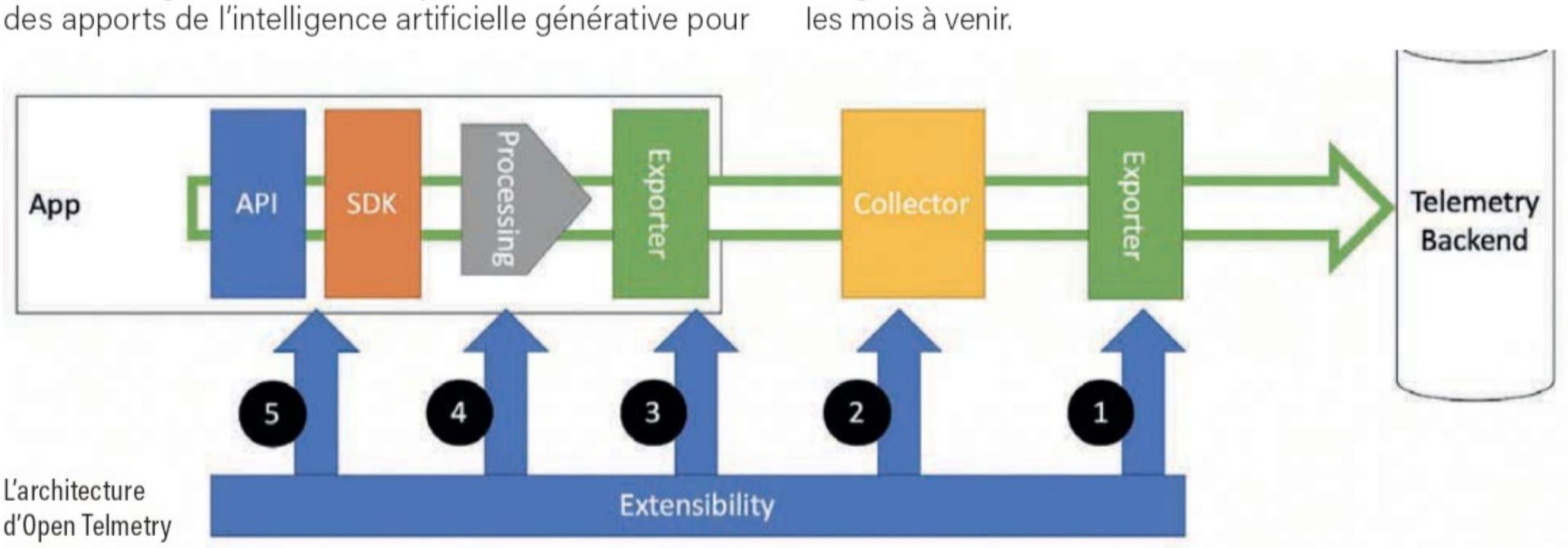
Autre approche, celle de Riverbed : *« on regarde le sujet du point de vue de l'utilisateur. Donc, l'expérience qu'il va avoir, la performance qu'il va assurer pour le business, le business qui va qui va diminuer ses coûts et améliorer ses revenus, ainsi de suite. Tout commence dans le poste de travail avec l'utilisateur quand il va cliquer quelque part. Donc, il va par exemple cliquer pour sortir une liste de clients ou bien cliquer pour sortir une liste de son inventaire qui est en stock, ainsi de suite. D'abord, ça commence par le poste de travail où il va faire ce clic pour demander, je veux la liste des clients ou bien je veux mon inventaire. Cette requête, la deuxième étape, elle va passer dans un réseau, dans un réseau WAN, dans un réseau LAN, pour arriver après à l'application. La deuxième partie après le poste de travail qu'il faut surveiller pour avoir une bonne observabilité, c'est la partie réseau. Donc, l'accès à l'application. Une fois qu'on arrive à l'application, il y a la partie Application Performance Monitoring avec l'analyse de code, ainsi de suite, pour s'assurer que l'application ou bien la requête s'exécute comme il se doit dans l'application. Mais ça ne s'arrête pas ici car l'application est hébergée sur un serveur, donc il faut aussi surveiller l'infrastructure. Si je veux aujourd'hui assurer que mon application métier tourne bien, il faut avoir une vision globale. On ne peut pas parler d'observabilité globale juste si on regarde l'application. Il faut regarder l'infrastructure où l'application est installée, il faut regarder le réseau d'accès sur lequel on peut avoir accès à cette application et ensuite, il y a le poste de travail de l'employé qui l'utilise pour accéder à cette application. Aujourd'hui, c'est ça notre vision chez Riverbed. Notre vision, c'est de partir sur du Unified Observability. On dit qu'on fait du Digital Experience Management ou bien Digital Employee Experience sur le poste de travail qui va nous amener le côté observabilité sur le poste de travail »*, détaille Joseph Slameh, director Solutions Engineering chez Riverbed. □

B.G

Les apports de l'IA et d'Open Telemetry

L'AIOps s'enrichit des nouvelles technologies d'intelligence artificielle et devient omniprésente dans les outils. De plus, une nouvelle approche est soutenue avec l'avènement d'un standard de fait : Open Telemetry.

Du fait de différentes problématiques comme le volume des données à traiter, le manque de ressources spécialisées, les différents silos de données, les outils d'observabilité ont de plus en plus recours à l'intelligence artificielle sous toutes ses formes. Cela reste principalement de l'apprentissage machine. Mais différents logiciels sur le marché intègrent des fonctions prédictives et bénéficient des apports de l'intelligence artificielle générative pour simplifier la remédiation ou l'identification des incidents pour des personnels peu qualifiés. Elle sert principalement à des fonctions d'automatisation. Si les éditeurs reculent encore à la rendre totalement autonome, elle peut d'ores et déjà prendre des actions seules sur des processus simples et encadrés. Elle est désormais présente dans la majorité des outils sur le marché. Ceux qui ne l'ont pas encore intégrée vont certainement le faire dans les semaines ou les mois à venir.



L'architecture d'Open Telmetry

Une vue de l'outil de Datadog d'observabilité des modèles d'intelligence artificielle

Un standard de fait

L'introduction d'OpenTelemetry (OTel) représente une véritable révolution pour l'observabilité dans le monde IT, puisque ce nouveau standard ou « framework » open-source consiste en une collection d'APIs, SDKs et outils pour instrumenter, générer et traiter des données sur la performance venant des logs, des traces ou des métriques dans un format unique et unifié, et facile à consommer par les développeurs. Le standard est maintenu par la CNCF (Cloud Native Computing Foundation) et est en train de devenir la norme de mise à disposition des données d'observabilité dans le cloud-native.

OpenTelemetry permet une approche normalisée et plus intégrée par rapport aux outils de monitoring traditionnels. Alors que ces derniers nécessitent une mise en place manuelle en connaissance de l'infrastructure, OpenTelemetry permet une instrumentation automatisée et dynamique garantissant l'observabilité en temps réel. Cette nouvelle norme est aujourd'hui clé dans la mise en œuvre d'une plateforme d'observabilité moderne permettant de mieux anticiper les problèmes, d'identifier en temps réel les signaux issus de l'ensemble de la chaîne applicative et d'apporter de la proactivité. Grâce à la capture automatisée et en temps réel des métriques, traces et logs au sein des environnements dynamiques et hybrides, les équipes IT peuvent s'apercevoir d'une détérioration d'un service dès l'instant où il devient visible. Ils peuvent donc anticiper et résoudre les problèmes avant qu'ils n'impactent réellement les utilisateurs. Cela permet une gestion plus fine et agile des infrastructures, où l'identification de signaux faibles devient un levier essentiel pour une meilleure prise de décision. Avec l'introduction d'OpenTelemetry, la manière dont les systèmes en production sont gérés se change au fur et à mesure. Où OTel n'était d'abord utilisé que pour les applications cloud-natives, on commence maintenant à l'appliquer pour les systèmes plus classiques grâce à l'avantage indiscutable de l'utilisation d'une norme commune et ainsi mieux comprendre le fonctionnement des systèmes en temps réel — aussi dans leurs interactions — et optimiser leur utilisation. Dans l'univers de production, chaque signal capté par OpenTelemetry permet d'évaluer les performances globales et de prévenir les dysfonctionnements.

L'observabilité avec OpenTelemetry permet également d'identifier les usages réels des applications et des infrastructures. Cela implique de pouvoir calibrer précisément et les éléments des configurations et leurs interactions, que ce soit au niveau des applications ou fonctions elles-mêmes, celui du middleware ou encore sur le niveau des infrastructures. L'objectif est de toujours trouver le juste équilibre entre l'agilité et la résolution proactive des problèmes, avant même que ceux-ci ne deviennent critiques. Bien que les bénéfices d'OpenTelemetry soient évidents, son adoption à grande échelle au sein des entreprises IT n'est pas sans défis. En effet, la mise en œuvre de ce framework d'observabilité nécessite une bonne dose d'expertise technique, ainsi que des ressources dédiées pour gérer l'instrumentation et l'analyse des données.

La plupart des logiciels du marché se sont déjà convertis à OTel, comme Splunk, Cisco, et bien d'autres, ce qui en fait d'ores et déjà un standard de fait, même s'il ne fait pas tout et est plutôt adapté aux environnements nativement cloud ou DevOps. □

B.G

Intel Xeon 6 P-Cores

Intel passe la vitesse supérieure

Intel a dévoilé, sur le MWC, de nouvelles versions de son Xeon de sixième génération. Cheval de bataille d'Intel sur le marché des serveurs, le Xeon gagne en puissance et, sans surprise, Intel pousse son processeur haut de gamme sur le marché IA...

Faute de mieux.

Intel continue de décliner sa gamme Xeon 6, nom de code Granite Rapids. Lors du Mobile World Congress (MWC), l'américain a dévoilé une nouvelle série de Xeon 6, conçus pour les applications réseau et Edge. Ces puces se destinent aux opérateurs télécom, ainsi qu'aux datacenters. Cette annonce fait suite aux lancements qui avaient lieu l'été dernier et au dernier trimestre 2024, avec des Xeon dotés de cœurs E (Efficient-cores), les « Sierra Forest ». L'idée est de fournir aux hyperscalers des serveurs dotés d'un maximum de cœurs pour porter les instances et les conteneurs de leurs clients. Ces Xeon 6 peuvent aligner jusqu'à 288 cœurs par socket.

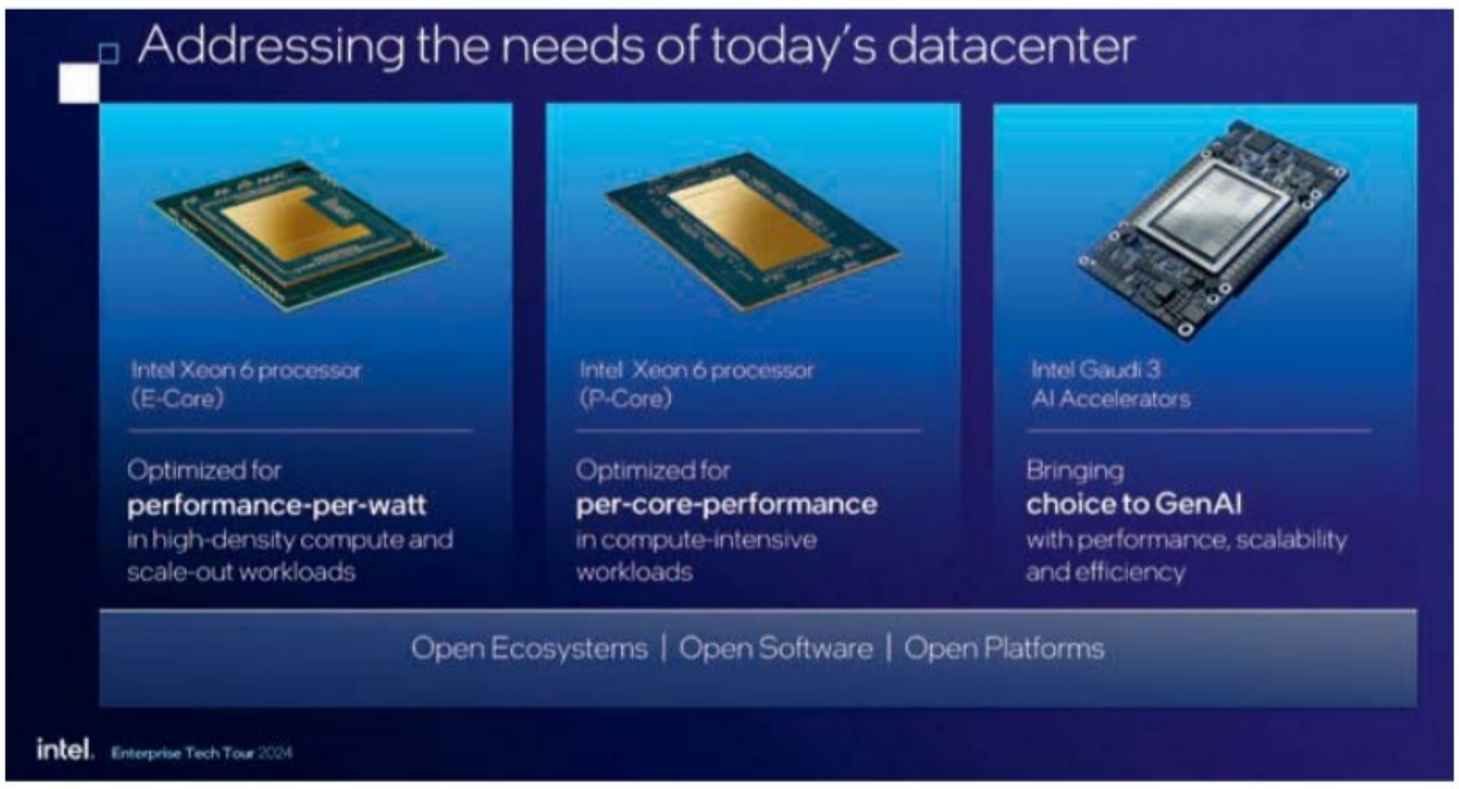
Des Xeon pour les opérateurs et le Edge

L'annonce Xeon 6 de Barcelone portait sur des SoC (System on Chip) dotés de cœurs plus puissants, les cœurs P (Performance-core). Ces composants sont destinés à exécuter des fonctions de routage des flux de la 5G, ce que l'on appelle le virtual Radio Access Network (vRAN), mais aussi permettre aux opérateurs de télécom de se positionner sur le marché du Edge Computing. En déployant des capacités de traitement en bas des tours 5G, ou dans les datacenters locaux, les opérateurs peuvent proposer une offre complémentaire aux datacenters géants et apporter de la puissance de calcul au plus près des utilisateurs. Bien évidemment, comme pour toute nouvelle génération de processeur, Intel annonce des performances bien supérieures à celles de la génération 5, une accélération d'un facteur 1,4 par rapport à la génération 5, d'un facteur 3,2 pour la performance IA, avec des processeurs comptant jusqu'à 128 cœurs. Au sortir du MWC, on était à 49 références, rien que pour la famille Xeon 6 P-cores !

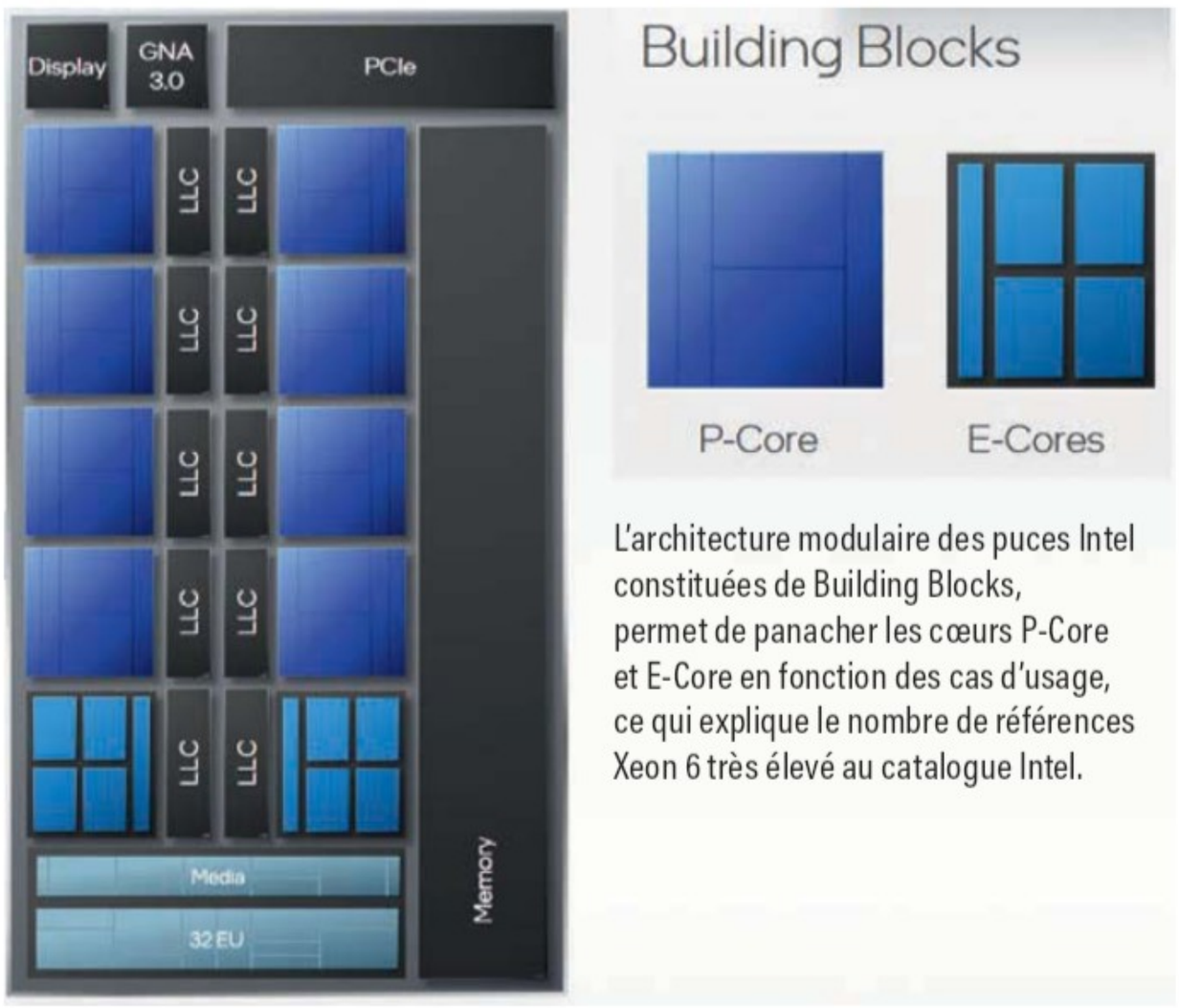
Si Intel multiplie les modèles, la logique reste assez simple : les puces dotées de P-Core dont les lancements se sont échelonnés en 2024 se destinent à l'IA, le HPC, les infrastructures, les puces dotées de E-Core iront à l'hébergement de sites web et d'applications. Pour le Edge Computing, les Powerpoint Intel laissaient penser à ces processeurs E-Core, mais finalement, l'annonce de Barcelone semble montrer que les P-Core ont finalement été choisis. En effet, Intel perd des parts de marché dans les datacenters et se retrouve complètement marginalisé sur l'eldorado du moment : les calculateurs dédiés à l'IA. Xeon 6 est envoyé dans la bataille.

Le cimetière des accélérateurs IA d'Intel

En 2019, Intel avait lâché 2 milliards de dollars pour prendre le contrôle de l'israélien Habana Labs et s'emparer du marché de l'entraînement des LLM avec la puce Gaudi. Amazon avait même sélectionné la puce pour ces datacenters... Finalement, le scénario ne s'est pas déroulé comme prévu et Intel n'a pu que grappiller quelques miettes du festin. En 2024, les ventes de Gaudi 3 ont été inférieures à 500 M$, un chiffre à comparer aux près de 61 Md$ engrangés par Nvidia sur la même période. Pire, l'architecture Falcon Shore qui devait succéder à Gaudi et permettre enfin à Intel de rebondir s'est avérée tellement catastrophique qu'Intel a dû jeter l'éponge. Le successeur de Gaudi a été abandonné avant même sa mise sur le marché. Un sacré coup dur après l'abandon des GPU Rialto Bridge en 2023.



La roadmap Intel du Xeon 6, présentée en 2024, est aujourd'hui tenue : les Xeon 6 à cœurs E et P sont au catalogue. Par contre, le volet GPU est un échec : Gaudi 3 ne tient pas ses promesses et son successeur est mort-né.

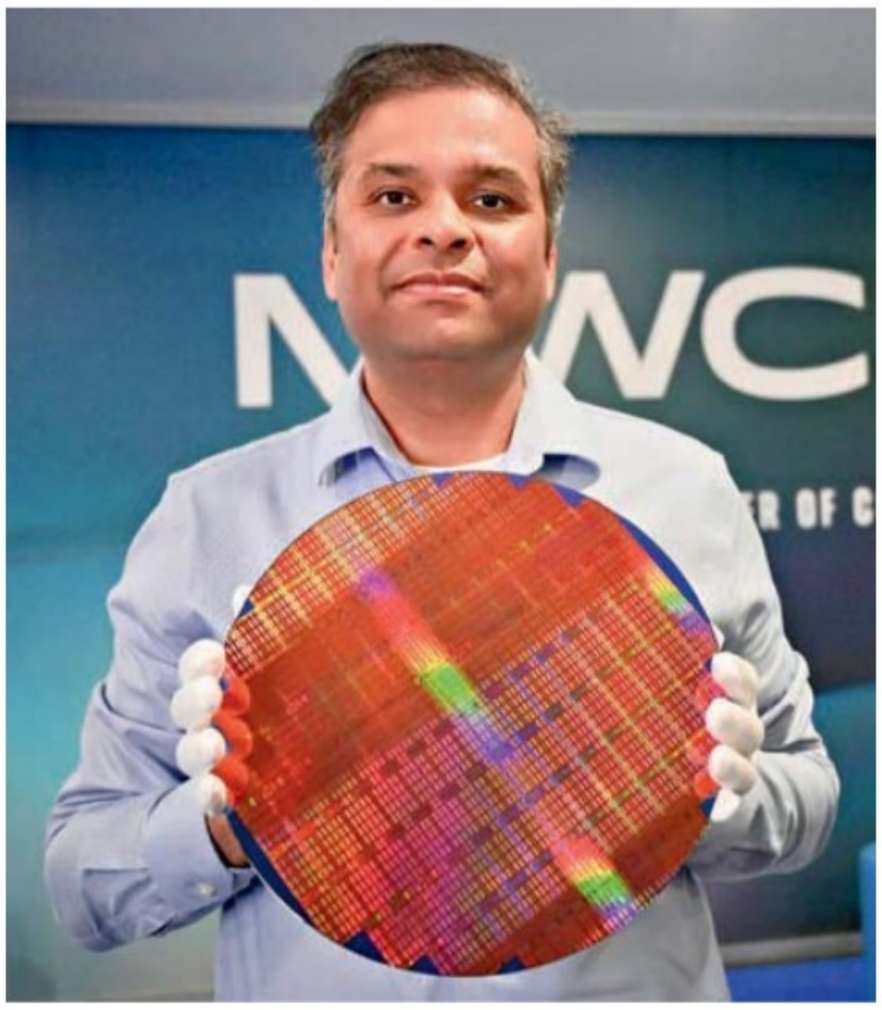


L'architecture modulaire des puces Intel constituées de Building Blocks, permet de panacher les cœurs P-Core et E-Core en fonction des cas d'usage, ce qui explique le nombre de références Xeon 6 très élevé au catalogue Intel.

Intel n'a plus rien à mettre sur ses étagères, mais ne peut se permettre de laisser le champ libre à Nvidia et un AMD qui retrouve des couleurs dans les datacenters. Les pertes colossales de l'industriel compliquent sérieusement la donne. Si Falcon Shore était considéré comme une simple évolution de Gaudi 3, Michelle Holthaus, la directrice produit qui assure l'intérim au poste de CEO après la mise à la retraite de Pat Gelsinger, a expliqué qu'Intel travaille sur la prochaine génération de GPU appelée Jaguar Shores. D'ici à ce que ce futur GPU arrive sur le marché, s'il arrive un jour, Intel n'a plus rien en catalogue pour contrer les H200 et Nvidia et les Instict MI300X d'AMD. Plus rien, sauf ce bon vieux Xeon et ses extensions IA…

De l'IA sur un Xeon, une idée qui remonte à 2019

Puisque la bataille est perdue face à Nvidia et AMD sur le marché de l'entraînement des IA, Intel se positionne sur celui de l'inférence en mode Edge. Le Xeon 6 ne va clairement pas lutter dans la même cour que les superchips (Assemblages CPU/GPU) de Nvidia ou d'AMD, mais, pour Intel, il peut porter les inférences d'IA dans les infrastructures Edge… L'idée n'est pas nouvelle. C'est en 2019, avec la génération Cascade Lake, qu'Intel a embarqué la technologie DL Boost (Deep Learning Boost) dans son processeur.



Sachin Katti, senior vice-président et directeur général Réseau et Edge Computing chez Intel, présente un wafer de Xeon 6 lors du MWC 2025 de Barcelone, le 2 mars 2025.

L'idée est alors de pousser le Xeon sur les plateformes dédiées à l'apprentissage et à l'inférence des IA. L'idée n'était pas si mauvaise, on était encore aux prémices de l'envol de Nvidia et tout était encore possible. La suite a montré que c'était une erreur. Les data scientists ont préféré l'écosystème hardware et logiciel de Nvidia, et le géant vert a grignoté des parts de marché sur les applications IA jusqu'en 2023, avant de réellement exploser à partir de début 2024.

En attendant la 18A…

Convaincre les entreprises de choisir le Xeon pour porter leurs inférences ne va pas être simple… De même, le futur de Xeon reste compliqué. Toujours confronté à des pertes financières abyssales, Intel a annoncé le report à 2026 des futurs Xeon Clearwater Forest produits en 18A. Comme l'expliquait Michelle Holthaus lors du CES en début d'année, 2025 sera une année charnière pour l'industriel. Intel a misé gros en sautant la génération 20A (20 angström / 2 nanomètres) pour aller directement au 18A et prendre de vitesse TSMC. La production en 18A de processeurs Panther Lake, des processeurs pour poste client, est censée avoir démarré au deuxième semestre 2024, mais ne devrait monter en volume que très progressivement. Beaucoup plus progressivement que prévu. Attendue pour 2025, elle ne pourrait avoir lieu qu'en 2026. Le rendement de la production oscillerait encore entre 20 et 30 %, un taux de rebut encore bien trop élevé pour élever la cadence.

Pour l'heure, TSMC assure la production en volume pour son partenaire, ce qui pèse lourdement dans les comptes d'Intel. Lorsque la production en 18A sera pleinement maîtrisée et que les Panther Lake seront produits en masse, alors viendront les Xeon Clearwater Forest et, peut-être, enfin un GPU pour l'IA digne de ce nom.

2025 s'annonce encore comme une année difficile pour Intel qui doit faire face à une lourde restructuration financière, faire monter en volume ses sites de production 18A et, accessoirement, se trouver un nouveau CEO. Pat Gelsinger avait tout misé sur le 18A pour relancer Intel. La réussite ou l'échec du ramp-up de Panther Lake va clairement donner une indication sur le futur d'Intel. □ **A.C**

Performance et sécurité

HPE présente la nouvelle génération de Proliant

Ces nouvelles machines s'appuient sur l'intelligence artificielle pour optimiser la gestion et veulent allier sécurité et performance.

Reposant sur une architecture à base de Xeon 6, les HPE ProLiant Compute Gen12 intègre la plateforme HPE Compute Ops Management, une solution cloud qui améliore la sécurité et l'efficacité énergétique des serveurs. Elle anticipe la consommation et permet de définir des seuils pour réduire les coûts et l'empreinte carbone. Une cartographie interactive facilite la gestion des infrastructures IT multisites. L'intégration d'outils tiers optimise la maintenance, évitant jusqu'à 4,8 heures d'indisponibilité par serveur chaque année selon HPE. Toutes ces fonctionnalités sont disponibles pour les serveurs HPE ProLiant Compute Gen12 et pour les générations précédentes à partir de Gen10.

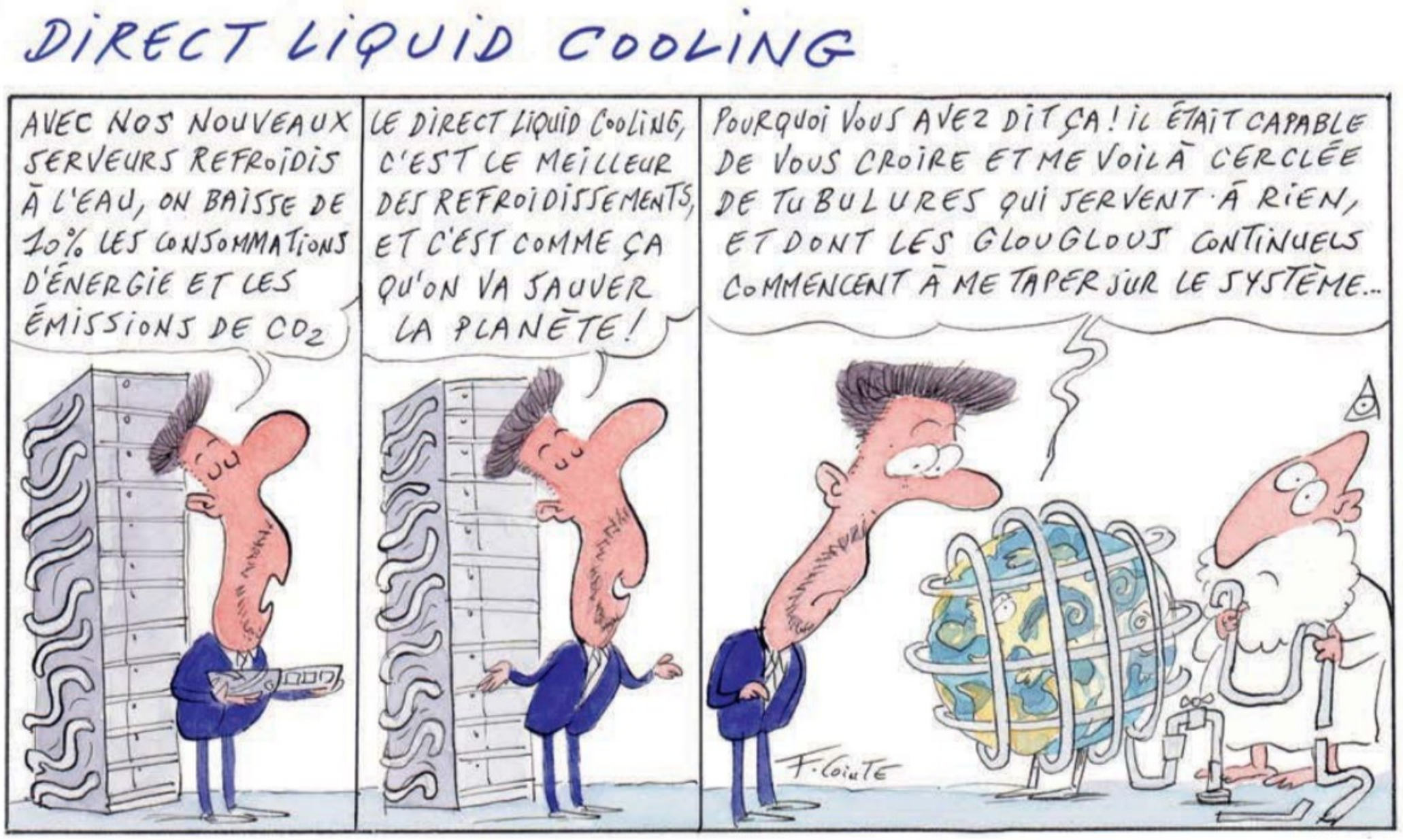
Une sécurité renforcée

HPE vise à proposer une sécurité *« du composant jusqu'au cloud »* tout au long du cycle de vie du serveur. Le système HPE Integrated Lights Out (iLO) 7 intègre désormais un processeur dédié à la sécurité : « secure enclave », propriété intellectuelle d'HPE améliorant la protection contre les attaques firmware. Les ProLiant intégrant l'iLO 7 offrent ainsi une protection avancée contre les cybermenaces, y compris celles liées à l'informatique quantique. Les nouveaux serveurs ProLiant respectent la norme cryptographique FIPS 140-3 niveau 3. HPE propose également un service sécurisé de mise hors service garantissant aux entreprises un recyclage conforme et la destruction des données en toute sécurité.

Des performances accrues

Selon le constructeur, les nouveaux serveurs sont optimisés pour des charges d'ampleur comme l'intelligence artificielle, l'analyse de données, l'edge computing, le cloud hybride et les infrastructures de bureaux virtuels (VDI). La performance est améliorée de 41 % par rapport à la génération précédente et il est possible de réaliser d'importantes économies de consommation électrique en consolidant jusqu'à sept serveurs Gen10 en un seul serveur Gen12. En option est proposé un refroidissement liquide sur toutes les nouvelles machines en rack. Six modèles de la gamme HPE ProLiant Compute Gen12 sont disponibles à la commande depuis le 24 février 2025, avec une livraison mondiale le 25 mars dernier. Ces modèles incluent les HPE ProLiant Compute DL320, DL340, DL360, DL380, DL380a et ML350 Gen12. D'autres modèles, comme le HPE Synergy 480 et le HPE ProLiant Compute DL580 Gen12 seront commercialisés à l'été 2025. L'ensemble des machines seront proposées via le portefeuille GreenLake. □

B.G



Stratégie

Lenovo mise sur le refroidissement liquide

Face à l'accélération des besoins en calculs liés à l'intelligence artificielle, Lenovo structure une stratégie autour de solutions datacenter refroidies à l'eau (DLC), capables de répondre aux problématiques énergétiques posées par l'IA. Il n'en oublie pas pour autant ses PC, qui devraient également prendre une partie de la charge.

« *Une machine IA de type DGX au format 8U va consommer entre 8 et 10 kW. Or, la plupart des salles en France ne sont tout simplement pas prêtes à délivrer une telle puissance* », alerte Cyril Fakiri, directeur technique Lenovo ISG France, à l'occasion du Lenovo Smart Innovation Tour qui s'est déroulé à Paris mi-mars. Au-delà de l'enjeu de la puissance de calcul, que la plupart des constructeurs sont aujourd'hui en mesure d'adresser via des partenariats avec Nvidia ou d'autres fondeurs, le fabricant chinois mise également sur les performances énergétiques de ses machines, notamment en matière de refroidissement. « *Dans les datacenters, le principal pôle de dépenses énergétiques reste le refroidissement* », tient ainsi à rappeler Cyril Fakiri.

Pour Jean-Christophe Morisseau, directeur général Lenovo ISG France, le refroidissement liquide devrait devenir la norme, au moins pour les charges de travail IA et le HPC.

Pour répondre à cet enjeu, Lenovo met en avant ses solutions de refroidissement à eau utilisant le Direct Liquid Cooling (DLC), développées depuis l'ère IBM, qui utilisait déjà le watercooling pour ses mainframes. « *Nous estimons qu'à plus ou moins long terme, ce mode de refroidissement deviendra la norme, au moins pour les charges de travail d'IA et le HPC. Le refroidissement liquide permet non seulement d'être plus économique en termes de consommation d'énergie, mais également d'améliorer la densité des infrastructures* », assure Jean-Christophe Morisseau, directeur général Lenovo ISG France.

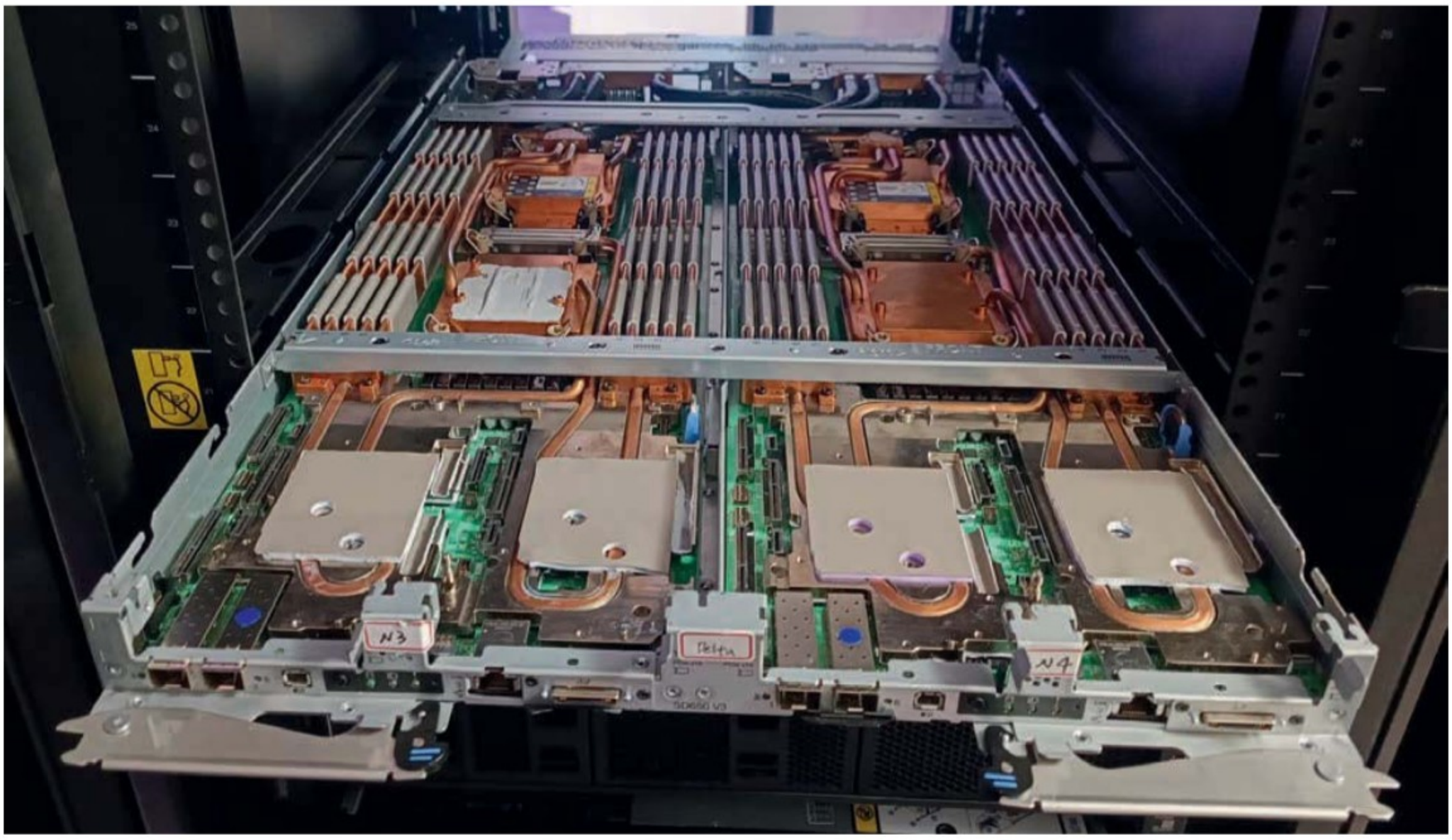
Des performances identiques, mais une consommation divisée par deux

Pour une infrastructure similaire à la DGX évoquée plus haut, le fabricant avance ainsi une consommation de seulement 5 kW, soit presque deux fois moins. Par ailleurs, les avantages ne se limitent pas uniquement à la consommation et à l'espace. « *Avec le DLC, nous avons un bien meilleur contrôle du refroidissement et, dans le cas de charges de travail IA ou HPC, il est possible de garder les composants à leur température optimale de fonctionnement et ainsi d'augmenter leur durée de vie. Au prix d'un GPU, ne pas pouvoir l'exploiter à son plein potentiel ou devoir même le couper à cause d'une surchauffe est plus que dommageable* », explique Cyril Fakiri.

Reste que cette technologie peut encore faire peur, notamment à cause des investissements initiaux et des changements d'habitudes qu'elle implique pour équiper les salles. « *Pour les solutions full DLC, il faut évidemment prévoir une arrivée d'eau et des solutions pour évacuer la chaleur fatale* », explique le directeur technique. Lenovo n'a d'ailleurs pas pu fournir d'indicateur de ROI précis entre le refroidissement par air et le refroidissement par eau. « *Quand la guerre russo-ukrainienne a éclaté, beaucoup de clients sont venus nous voir pour trouver des solutions moins gourmandes en énergie afin d'absorber la hausse des coûts. Tous ceux qui ont fait le choix du DLC en sont aujourd'hui ravis et continuent d'en déployer* », avance cependant Cyril Fakiri. « *Les spécialistes des datacenters comme Digital Realty et Equinix équipent aujourd'hui toutes leurs infrastructures avec des arrivées d'eau* », rappelle de son côté Jean-Christophe Morisseau.

Une offre modulaire pour monter en puissance sans se mouiller

En outre, le DLC peut également être déployé dans des salles « classiques » avec une montée en charge progressive. « *Avec la gamme Neptune Core, on va proposer des solutions sur des racks standards capables d'accepter progressivement du refroidissement liquide pour certains composants. L'idée est de pouvoir facilement combiner refroidissement par air et refroidissement par eau* », détaille Cyril Fakiri. La gamme de serveurs Neptune Air embarque pour sa part des systèmes de refroidissement liquide complets, avec un circuit fermé équipé de son propre échangeur pour chaque serveur, et peut donc aussi être déployée sans arrivée d'eau.



Les solutions DLC de Lenovo embarque un circuit d'eau qui vient refroidir chaque composant du serveur.

Dans un contexte où les salles informatiques sont vieillissantes (avec un PUE de 1,8 ou plus), Lenovo propose également des alternatives avec ses partenaires. ModulIT a ainsi déployé une offre de conteneurs modulaires, prêts à l'emploi, embarquant des infrastructures refroidies à l'air ou à l'eau. Ces micro-datacenters, totalement autonomes, peuvent être déployés temporairement en attendant une montée en charge des salles principales, ou en mode pérenne dans des environnements contraints.

Une chaîne du edge au datacenter

L'autre force de Lenovo est d'intégrer sa stratégie IA de bout en bout, dite du « pocket to cloud ». Avec ses PC IA dotés de NPU compatibles Copilot+, le fabricant s'est positionné sur plusieurs projets de computer vision pour la smart city, la surveillance ou l'industrie. Il propose également des infrastructures edge prêtes à l'emploi pour couvrir tous les maillons de la chaîne. Le constructeur mise sur une approche écosystémique avec ses partenaires pour proposer des bundles complets (caméras, logiciels, IA embarquée, etc.).

Avec l'IA qui bouleverse les architectures IT, Lenovo met donc en avant le savoir-faire hérité d'IBM en matière de refroidissement liquide pour proposer une alternative viable, réaliste et modulaire. Son offre Neptune, que le fabricant aimerait bien voir devenir une architecture de référence pour les charges de travail les plus gourmandes, permet, grâce au refroidissement liquide, d'augmenter la densité des infrastructures tout en les rendant moins énergivores. Sa stratégie de « pocket-to-cloud », supportée par un stockage qui va enfin offrir des solutions haute performance dignes de ce nom (voir encadré) doit également lui permettre de s'imposer sur l'IA. □

O.B

STOCKAGE — LE PARENT PAUVRE QUI NE LE SERA BIENTÔT PLUS

***« Nous sommes en train de monter en puissance sur le stockage pour proposer enfin des offres high-end pertinentes »*, assure aujourd'hui Jean-Christophe Morisseau, qui rappelle que Lenovo est longtemps resté un fabricant OEM de NetApp. Afin de s'affranchir des limites d'ONTAP, Lenovo a d'abord commencé à nouer des partenariats avec d'autres fournisseurs, notamment Vast Data et Weka, afin de proposer des solutions de stockage haute performance. Le grand pas en avant devrait toutefois arriver dans le courant de l'année avec la finalisation de l'acquisition d'Infinidat.**

L'opération, qui devrait être bouclée d'ici la fin du printemps, va permettre à Lenovo d'intégrer à ses offres les solutions de stockage haute performance de l'entreprise israélienne. Cette dernière est spécialisée dans les baies hybrides combinant composants flash et disques durs, associées à une plateforme logicielle avancée en mode SDS. *« L'objectif est d'intégrer de l'IP en propre dans nos baies et de proposer une offre cohérente du calcul au stockage, capable de soutenir les traitements lourds liés à l'entraînement des LLM et à l'inférence »*, déclare ainsi Jean-Christophe Morisseau.

L'IA, moteur d'une révolution silencieuse dans les data centers

Avis d'expert de Jean-Pierre Tournemaine, Directeur Régional pour l'Europe du Sud chez Vertiv, fournisseur mondial de solutions de continuité et d'infrastructures numériques critiques

En 2025, l'intelligence artificielle (IA) s'impose comme le catalyseur d'une expansion du secteur des data centers. Cette évolution, visible du grand public, offre des opportunités majeures à l'économie numérique et présente de nombreux défis liés à l'efficacité énergétique. Les experts de Vertiv anticipent une année marquée par l'innovation.

Une densification sans précédent

L'IA redéfinit les standards de l'infrastructure numérique. Les racks traditionnels, qui supportaient en moyenne 8,2 kW en 2020, devraient désormais gérer des puissances de 200 à 500 kW. L'association professionnelle France Datacenter estime également que la consommation moyenne d'un data center devrait augmenter de 64% entre 2024 et 2026[1]. Cette densification impose une refonte des systèmes d'alimentation et de refroidissement. Le refroidissement liquide, notamment en direct-to-chip ou immersion, s'impose comme des solutions incontournables pour gérer ces charges extrêmes (au-delà de 30 à 40 kW par rack). Les architectures hybrides (refroidissement par air et « liquid cooling ready »), combinant différentes technologies, se développent pour s'adapter à tous types d'environnements.

Les data centers s'équipent également de systèmes d'alimentation sans interruption (ASI) et d'équipements de distribution électrique capables de gérer des fluctuations de charge importantes, typiques des workloads d'IA (de 10 % à 150 % de surcharge instantanément).

Le défi énergétique et la durabilité

L'essor de l'IA accentue la pression sur les réseaux électriques. La part des data centers dans la consommation énergétique mondiale devrait passer de 1-2% actuellement à 3-4% d'ici 2030. En France, RTE prévoit ainsi que la consommation des data centers français devrait atteindre entre 23 et 28 TWh en 2035, soit environ 4% de la consommation électrique nationale[2]. Cette augmentation soulève des questions de durabilité et d'approvisionnement. Le secteur s'oriente donc vers le déploiement accéléré de micro-réseaux intelligents ainsi que l'exploration de technologies et d'énergies alternatives. En réponse à ces défis, des solutions comme le refroidissement liquide de Vertiv contribuent aux initiatives d'économie circulaire en permettant une réutilisation efficace de la chaleur des centres de données. Leur technologie capture la chaleur résiduelle générée par les équipements informatiques grâce à des systèmes de refroidissement liquide, qui peut ensuite être réutilisée pour le chauffage urbain ou d'autres processus industriels. En mettant en œuvre une telle infrastructure de refroidissement liquide, les centres de données peuvent transférer l'énergie thermique capturée vers des installations ou des communautés voisines, transformant ce qui serait autrement de la chaleur fatale en une ressource précieuse. Cette approche permet non seulement de réduire la consommation d'énergie globale et l'empreinte carbone des centres de données, mais aussi de promouvoir les pratiques durables en créant un système en boucle fermée où la chaleur fatale devient un élément utile pour d'autres applications, démontrant ainsi une mise en œuvre pratique des principes d'économie circulaire dans les opérations des centres de données.

Face à ces défis, une collaboration s'installe entre les acteurs de l'écosystème : développeurs, fabricants de semi-conducteurs et de solutions d'infrastructure, fournisseurs d'énergie. Cette synergie vise à soutenir l'adoption de l'IA et accélérer l'innovation.

2025 s'annonce comme une nouvelle année prometteuse pour le secteur des data centers. L'IA, tout en offrant des opportunités sans précédent, impose une transformation du secteur. La capacité à innover, à collaborer et à s'adapter aux nouvelles réglementations sera cruciale pour relever ces défis et soutenir un développement plus durable des infrastructures numériques. Les acteurs du secteur devront naviguer entre innovation technologique, efficacité énergétique et réglementation, tout en répondant aux exigences croissantes en matière de performance et de sécurité. L'avenir des data centers se dessine ainsi à travers une approche holistique, où technologie, durabilité et régulation fusionnent pour façonner l'infrastructure numérique de demain.

1 : France Datacenter 2024 Étude d'impact économique, social et environnemental de la filière des datacenters en France Présentation de l'étude Juin 2024

2 : Transitions, Le Magazine de RTE qui éclaire les futurs énergétiques — n°11 — Décembre 2024

Salaires

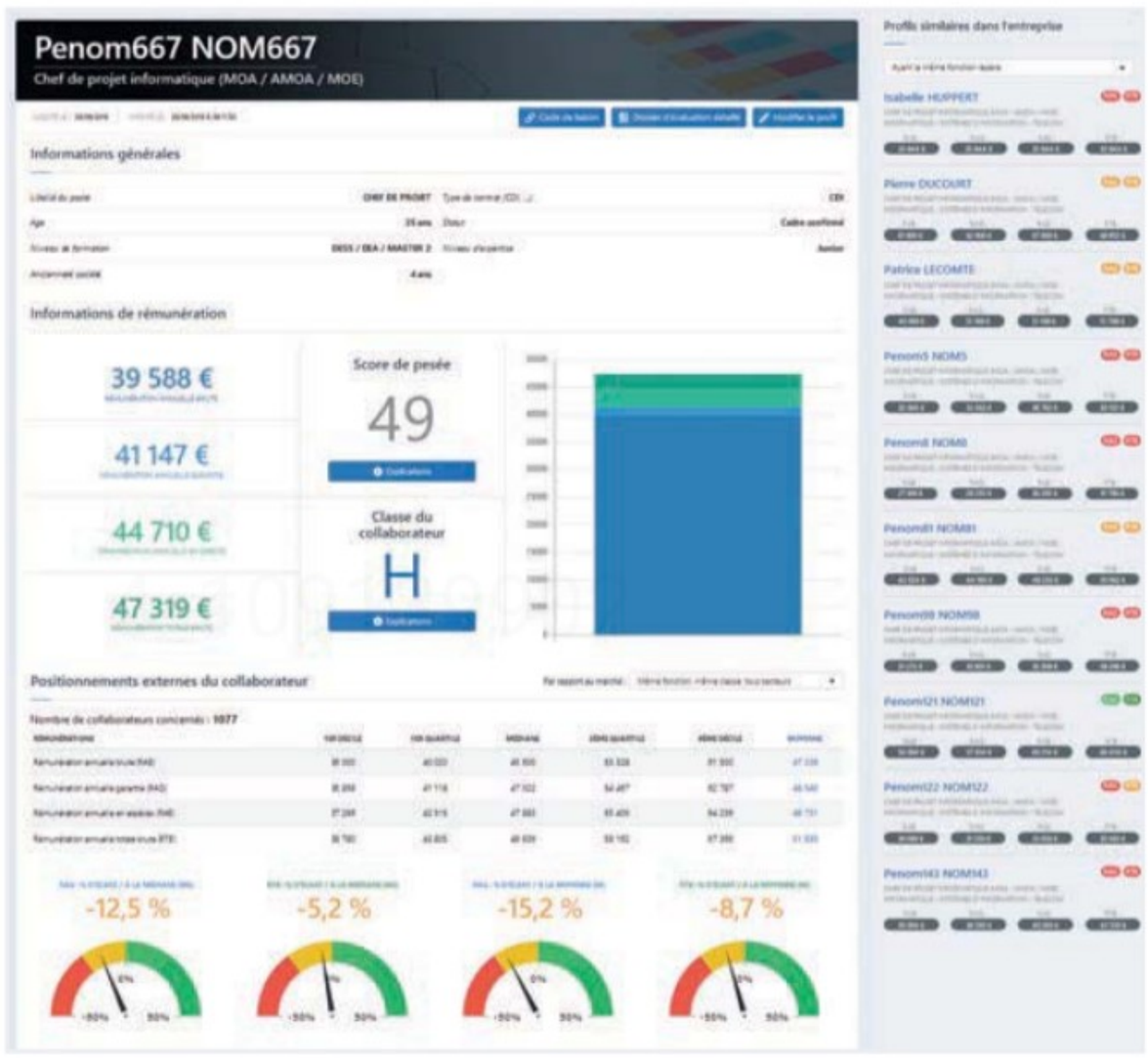
People Base face aux cabinets américains

People Base CBM (Compensations and Benefits Management) est un cabinet de conseil français et indépendant, spécialisé en stratégie de rémunération et politique salariale des entreprises. Dans un contexte où la souveraineté numérique devient un enjeu majeur, People Base CBM s'impose comme l'alternative française conforme aux exigences réglementaires européennes.

Spécialiste français du conseil en stratégie de rémunération, People Base CBM accompagne les entreprises dans la gestion et l'optimisation de leurs politiques salariales, avec une approche souveraine et 100 % conforme aux réglementations locales. En proposant des solutions développées et hébergées en France, People Base CBM garantit aux entreprises la maîtrise de leurs données stratégiques, tout en leur offrant une expertise de haut niveau.

Le cabinet aide les entreprises à bâtir des politiques de rémunération alignées sur leurs enjeux sectoriels, leur positionnement concurrentiel et leurs objectifs de croissance. Contrairement aux approches standardisées des cabinets internationaux, People Base CBM propose des solutions sur mesure, conçues pour maximiser l'attractivité et la performance des entreprises françaises.

Une vue de Waage Pro



Une plate-forme en SaaS

Dans cette démarche, People Base CBM a conçu Waage Pro, une plateforme de gestion des rémunérations totalement indépendante des infrastructures étrangères. Conçue, développée et hébergée en France, elle permet aux entreprises de comparer leurs politiques salariales aux standards du marché, de gérer leur masse salariale avec précision et d'optimiser leurs décisions stratégiques en toute sécurité. Par une interface web, Waage Pro apporte une suite d'outils destinés à aider les entreprises dans la gestion de leurs ressources humaines, et plus spécifiquement dans le pilotage des rémunérations et de leur politique salariale (bases de données rémunérations, benchmark et enquêtes de rémunérations, outils de gestion et de pilotages RH, logiciels d'administration RH…). La solution comprend une base de données des rémunérations dans plus de 500 postes. Le module de benchmark de rémunérations permet de connaître en temps réel le positionnement salarial de chacun de ses collaborateurs par rapport à plusieurs marchés de références. La plateforme permet, de plus, de détecter les situations critiques, afin de prévenir la fuite de collaborateurs clés pour l'entreprise, ainsi que la réalisation d'analyses poussées sur les pratiques salariales de la société et sur sa masse salariale : analyses générales des rémunérations, analyses des niveaux de rémunérations proposés, analyses sur la structure des rémunérations versées, analyses sur la parité professionnelle… Le module « executive » a été créé pour suivre, analyser et positionner avec une grande précision, les rémunérations des dirigeants d'une entreprise. Il permet en effet de consulter les pratiques du marché (données quantitatives et qualitatives), obtenir les positionnements automatiques ou sur mesure des dirigeants (dirigeants mandataires sociaux ou non et membres du comité de direction : directeur financier, directeur des ressources humaines, directeur marketing, directeur informatique, de la production, etc.), et de visualiser les principaux organigrammes évalués de l'équipe de direction. Waage Pro calcule automatiquement, à partir des informations saisies dans le portail, le « score général » de chaque société abonnée au portail. Ce score général est lui-même calculé à partir de quatre critères majeurs que sont la compétitivité de l'entreprise, l'équilibre interne des rémunérations, le respect de la parité professionnelle et la pertinence de la politique de rémunération. □

B.G

Service public

Inetum Software devient Nexpublica

La branche software du groupe inetum reprend sa liberté et change de nom pour revenir à son cœur de métier, les logiciels pour le service public.

Martin Hubert, président de Nexpublica explique : *« Ce qu'on a décidé, c'est de revenir à nos premières amours, donc le secteur public, très récemment. Dans la réalité, les liens entre nos activités de software et les activités de services du groupe Inetum sont ténus. Et donc, on parlait depuis longtemps de cette possible séparation. Nous l'avons fait pour avoir une autonomie stratégique, finalement précieuse pour la suite de notre activité, avec des investissements R&D plus forts et de potentielles acquisitions, pour devenir le leader incontournable de notre marché, le logiciel pour le secteur public. »*

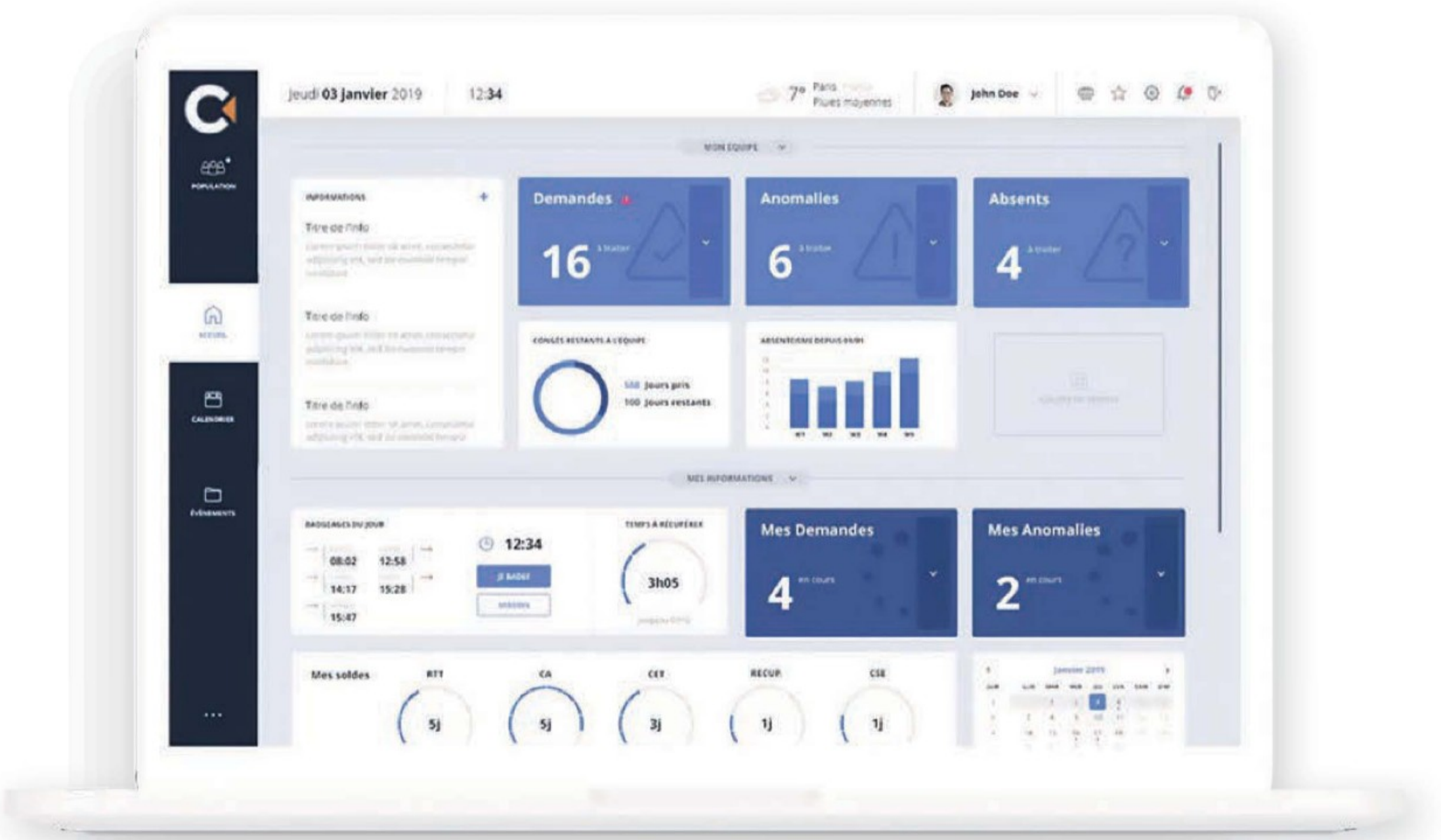
Traiter les différents cas d'usage

Le dirigeant ajoute : *« selon les domaines, nous pouvons avoir des logiciels qui diffèrent selon les cibles. Entre les collectivités locales, de très grande taille ou de très petite taille, entre les établissements publics, les ministères ou les entreprises parapubliques, on peut avoir des sujets qui diffèrent. Par contre, il y a des sujets communs. Exemple, la paye du secteur public est un sujet commun, sauf pour les ministères. Mais après, pour certains sujets, par exemple les aspects financiers, on peut avoir des sujets qui diffèrent. Nous avons des logiciels qui couvrent une très grande partie du secteur public, une grande partie de ses cibles, mais pas encore toutes les cibles et pour tous les domaines ».* Il continue : *« nous avons besoin d'intégrer l'intelligence artificielle, de faire ce que le ministère de la fonction publique avait appelé le service public augmenté autour de chacune de nos priorités sur l'attractivité du métier d'agent public ou la proximité du secteur public et du citoyen. C'est un exemple des domaines fonctionnels que nous devons ajouter à nos produits existants, par exemple autour de la gestion des déchets, ainsi que de nouveaux domaines que nous ne couvrons pas aujourd'hui, que des acteurs sur le marché couvrent bien et par lesquels nous voulons réaliser des acquisitions externes pour compléter notre portefeuille. Le domaine des RH dans la fonction publique est un domaine qui nous intéresse, on est en train de lancer la suite RH la plus avancée pour le secteur public, en combinant nos forces autour de la gestion des temps et des activités et autour de la paye, de la gestion des talents. On peut encore compléter notre offre par d'autres composants fonctionnels, cela fait partie des sujets qui nous intéressent, mais la gestion administrative, les services techniques, les services aux citoyens nous intéressent aussi pour compléter notre portefeuille. »*


© Thomas Laisne - La Company

En SaaS sauf...

Le dirigeant indique que tous les logiciels sont « Cloud ready », mais que certains clients ou administrations souhaitent avoir les produits sur site ou dans des clouds privés déployés en interne (pompiers, SAMU...). L'ensemble des offres Ccoud sera rassemblé sur celui de S3NS. Martin Hubert commente : *« le côté souverain est bien entendu clé pour ce qui nous concerne et pour le secteur public, pour nos*



Une vue du logiciel Chronotime de Nexpublica

clients. Pour le Secnum Cloud, nous y allons à la vitesse du secteur public. Plus le secteur public voudra faire du Secnum Cloud, plus nous serons là pour les accompagner et même anticiper avec eux ce sujet. Et c'est ce sur quoi nous travaillons avec S3NS dès à présent ».

Aider à la transformation du secteur public

Interrogé sur ces priorités, Martin Hubert ne se dérobe pas : *« Notre priorité, c'est dans un premier temps, d'être un partenaire technique du secteur public qui aide à prendre de la hauteur. Nous sommes un partenaire du secteur public depuis plusieurs dizaines d'années. La première chose consistera donc, d'ici à quelques semaines, à publier un manifeste, une forme de vision de l'évolution du digital dans le secteur public, sur les différents thèmes d'enjeux du secteur, c'est-à-dire la proximité du citoyen et du secteur public, l'attractivité du métier d'agent. Il y a eu 11% d'offres en moins sur les six dernières années dans le secteur public, mais il y a eu 50% de candidats en moins. C'est la cybersécurité et la souveraineté, par exemple. Prendre de la hauteur sur ces sujets, avoir exprimé une vision claire, et la décliner sous la forme de R&D, de roadmap et de démarches d'approche de nos clients du secteur public sur le sujet, c'est ça notre enjeu. »*

Rapprocher les citoyens

« La proximité entre le citoyen, l'usager d'un côté et le service public de l'autre est notre premier enjeu. Pour ce sujet, nous travaillons par exemple sur la mobilité, de façon à pouvoir accéder à distance au plus grand nombre de services publics. Nous travaillons avec l'IA sur des sujets comme regarder la validité des pièces, des dossiers qui sont remis dans le domaine social, dans le domaine du transport scolaire, etc. De façon à pouvoir valider à distance les pièces le plus simplement possible, avec une expérience usagée qui soit améliorée, pour simplifier la vie du citoyen. C'est d'ailleurs le sujet-même de notre remarque, puisque Nexpublica signifie nouvelle expérience publique et ce, pour tous les publics. »

Le dirigeant estime : *« On est déjà très fortement impliqués sur ces roadmaps, puisque c'est un projet qui a démarré depuis quelques mois déjà. On peut considérer que d'ici 24 mois, nous aurons déjà très significativement atteint les premiers objectifs que nous nous sommes assignés. »*

Une internationalisation à étendre

Le développement international est aussi sur la feuille de route de Nexpublica. Martin Hubert précise : *« on est présent dans un certain nombre de pays, en Espagne, en Suisse, en Pologne, au Luxembourg, par exemple, mais nous souhaitons dans un premier temps être le leader incontournable de notre marché en France, et progressivement internationaliser de plus en plus notre gamme de produits. Cela se fera, selon les sujets, soit par des développements internes, en réutilisant les composants logiciels que nous avons en France, soit par des acquisitions externes, lorsque l'adhérence réglementaire est trop forte et trop différente de la France. »* □

B.G

La revanche de l'open source

par Bertrand Garé

La Kubecon, la conférence de la CNCF sur Kubernetes, et sur l'actualité de la fondation vient de s'ouvrir à Londres. Cela nous a donné l'idée de revenir sur le parcours de l'open source au cours des dernières années. Qui se rappelle que les débuts ont été difficiles, concentrés sur des guerres de chapelles entre logiciels propriétaires forts de leurs années de développement et de nouveaux logiciels évidemment moins aboutis et loin de ce que l'on pouvait appeler le standard des entreprises. Puis peu à peu, l'open source a conquis les infrastructures quasi subrepticement. Elle est désormais présente du serveur au stockage, aux applications, elle remet aujourd'hui en cause un monde bien installé dans la virtualisation vers une pile complète s'appuyant sur Kubernetes et les containers. Mieux, Microsoft est dédormais le sponsor Diamant de la Kubecon. Comme quoi l'industrie IT change d'avis aussi vite que ses intérêts financiers.

Toutes les premières interrogations et contraintes autour de l'open source aujourd'hui sont levées. D'ailleurs, il devient rare de voir une nouvelle entreprise ne pas se prévaloir d'être open source et de donner son logiciel à une communauté chargée de renforcer les forces de développement en place dans l'entreprise qui propose le logiciel. Chris Anyczczyk, CTO de la CNCF (cloud native computing foundation) a indiqué lors de sa session plénière que le cloud native avait atteint un niveau d'adoption de 89 %. Selon un sondage réalisé pour le compte de la fondation, 93 % des entreprises interrogées indiquent utiliser ou évaluer Kubernetes pour le pilotage, renforçant ainsi la stature de standard de fait de Kubernetes dans les infrastructures des entreprises.

L'innovation, d'où qu'elle vienne est donc aujourd'hui open source. C'est la première revanche de l'open source sur le monde dit « propriétaire ». Cela fait longtemps que celui-ci s'est rallié aussi à l'open source pour alimenter son pipeline de nouveautés. Cela pose cependant question autour de ce qu'est une communauté open source aujourd'hui. Des développeurs « garage » des débuts, l'open source est aujourd'hui à la tête d'une armée de salariés d'entreprises qui développent des projets open source. Ceux-ci développent cependant en vertu des intérêts des entreprises qui les emploient. Certains projets sont mêmes associés plus ou moins à une seule entreprise qui devient ainsi le principal contributeur, mais aussi dirige les développements futurs du logiciel en fixant la feuille de route agrémentée de quelques autres projets connexes que la communauté se charge de mettre au niveau du logiciel principal.

De moins en moins de fondation

Selon les besoins et l'actualité des projets, on constate une érosion du nombre de fondations consacrés au développement de projets open source. Les évolutions de l'industrie font que certains projets tombent en désuétude, remplacés par une nouvelle vague de projets. Ainsi, Cloud Stack et Open Stack, pardon Open Compute, ont rallié une fondation plus puissante. Il en est de même pour Ceph, dans le stockage. Ces nouveaux mastodontes en charge de nombreux projets, Linux fondation, CNCF et d'autres rythment maintenant la vitesse des évolutions dans les entreprises.

L'open source est reconnue par le business

Une étude de Serena, rendue publique le 10 avril, a analysé les données sur l'historique de l'open source commercial. Elle indique que les investisseurs n'hésitent

plus à mettre de l'argent dans des entreprises prônant l'open source, quand elles sont dans les secteurs qui composent la colonne vertébrale de l'écosystème du cloud et du développement moderne de logiciel. Infrastructure, IA, sécurité, outils de développement, applications métiers et blockchain sont ceux dont les poches se remplissent avec de l'argent frais des venture capitalists. L'étude démontre ainsi que les entreprises adoptant un modèle d'open source commercial lèvent plus facilement et plus rapidement des fonds que des entreprises au modèle commercial classique. Elle constate de plus que le montant des fonds levés sont de plus en plus importants. Ce type d'entreprises peut même assurer de nouveaux tours de financement malgré une faible communauté suivant le projet. D'ailleurs, l'importance de la communauté devient de moins en moins importante, prenant en compte que les projets sont maintenant réellement conduits par les salariés de l'entreprise et non plus par des développeurs indépendants extérieurs à l'entreprise conduisant le projet. Le poids de la communauté ne pèse réellement dans le financement que lors de la phase d'incubation, et ne présage pas réellement des sommes levées dans les tours suivants. 12 % des entreprises ayant adopté ce modèle d'open source commercial ont réussi leur sortie que ce soit par une entrée en bourse ou par une fusion-acquisition. Sur les 850 entreprises recensées par l'étude, 110 sur les 25 dernières années ont ainsi effectué ce parcours : 24 sont entrées en bourse et 86 ont fusionné. Le parcours moyen de ces entreprises est de quatre ans et demi avec un financement de 47 M$ pour des valorisations dépassant les 400 millions de dollars. Talend, Cloudera, Mulesoft, Rackspace en sont des exemples éclairants.

Rien que par ces éléments, l'open source prend une revanche éclatante sur les jugements des éditeurs propriétaires du début de sa vie. Ils sont désormais tous rangés sous sa bannière. D'où peut-être leur volonté de conduire désormais les projets en devenant les sponsors premium et en apportant de larges contributions, afin de conduire la feuille de route de celui-ci, si besoin, pour préserver leur intérêt sous le couvert de l'innovation et des « demandes des clients ». Comme si, du jour au lendemain, les clients se réveillaient pour avoir ceci ou cela. L'industrie informatique a toujours été une industrie de l'offre, trop rarement de la demande et l'open source n'a pas réussi à changer cela, c'est peut-être le revers de la médaille, le seul échec de cette manière de concevoir l'informatique. En effet, aujourd'hui nous sommes loin des débats entre Richard Stallman et Microsoft. L'open source règne. Longue vie à l'open source ! □

Direct-to-Cell

La menace Musk pèse sur la téléphonie mobile

On l'appelle le « Direct-to-Cell », et ce service pourrait bien bousculer le monde des télécoms. Imaginer envoyer des SMS même là où il n'y a aucune couverture cellulaire, c'est maintenant possible. Les premiers services émergent. Apple avait montré la voie avec ses messages d'urgence. Elon Musk et SpaceX débarquent avec T-Mobile aux USA.

Le 26 novembre dernier, la FCC, autorité américaine en charge des télécoms, donnait son accord à SpaceX pour opérer des services de type Direct-to-Cell aux abonnés de T-Mobile aux États-Unis, à partir de sa constellation de satellites déjà en orbite. Quelques jours plus tard, T-Mobile lançait les inscriptions à la beta du service. L'opérateur présente cette offre comme le moyen d'envoyer des messages texte de n'importe où aux États-Unis, y compris les 1,3 million de km² de zones blanches non couverts par ses antennes. Ce service sera ensuite intégré sans surcoût au forfait Go5G Next, le haut de gamme de l'opérateur. Les abonnés aux autres forfaits y auront accès pour 5 dollars par mois, puis 15 dollars par mois à partir du mois de mars. Le service démarre son activité avec les SMS seulement, mais promet les transferts d'images et les appels vocaux pour plus tard…

Avec son service d'appels de secours par satellite, Apple a été précurseur dans le Direct-to-Cell. Starlink pourrait bien banaliser l'approche à tous les smartphones.

De nombreux smartphones déjà compatibles

Si Bloomberg a révélé qu'Apple travaillait secrètement avec SpaceX depuis des mois pour implémenter cette compatibilité Starlink dans iOS, la liste de smartphones compatibles publiée par T-Mobile est déjà plutôt longue. Tous les modèles datant de moins de 4 ans sont potentiellement éligibles à la beta. Depuis les iPhone 14, les Google Pixel 9, Motorola 2024 aux Samsung Galaxy A14, S21, X Cover6 Pro, Z Flip3 et Z fold3, en ajoutant REVVL 7, des millions de devices vont pouvoir converser avec l'espace.

Cette compatibilité avec les iPhone est une surprise, car Apple proposait déjà un service Direct-to-Cell monté avec Globalstar, un opérateur de satellites dans lequel il a injecté 1,1 milliard de dollars fin 2024. C'est lui qui assure la fonction *« SOS d'urgence par satellite »* disponible aux États-Unis et au Canada depuis iOS 16.4. Le lancement de cette fonction avait même poussé Qualcomm à signer un partenariat avec Iridium lors du CES 2023. Partenariat finalement abandonné, faute d'intérêt par les constructeurs.

Comme souvent avec les promesses d'Elon Musk, celles-ci se concrétisent avec retard… C'est le 25 août 2022, que le chef ingénieur de SpaceX et Mike Sievert, CEO de T-Mobile, annonçaient depuis la Starbase le projet Coverage Above & Beyond.

Des SMS et des SOS uniquement, pour l'instant !

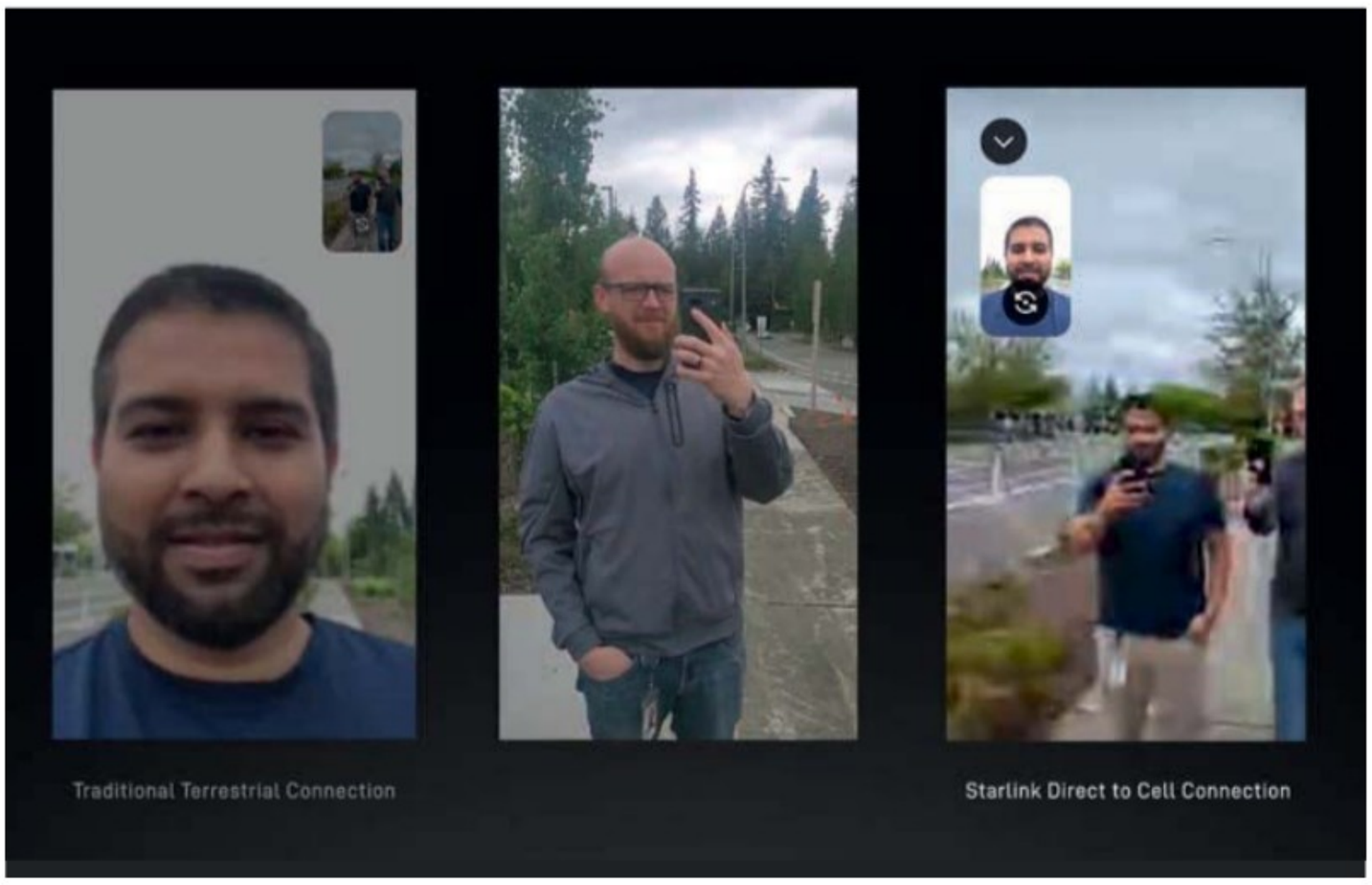
Pour l'heure, la FCC estime qu'un tel service permettra de donner accès au service de secours aux personnes qui n'ont pas de couverture mobile. Néanmoins, l'autorité américaine souhaite évaluer les éventuelles interférences que pourrait produire ce service, notamment vis-à-vis des autres opérateurs de satellites. Elle donnera alors son feu vert à un accroissement de la puissance d'émission des satellites Starlink, afin de transmettre autre chose que du texte.

Pour l'instant, toute la constellation Starlink ne supporte pas le Direct-to-Cell, révèle Renaud Kayanakis, associate partner chez Sia Partners : *« Seule une partie des satellites Starlink en orbite sont compatibles avec le Direct-to-Cell, ce sont les modèles Starlink Block v2-Mini-D2C qui sont deux fois plus gros que la génération précédente. Selon le site T-Systems, il est indiqué que 477 satellites sont compatibles, un nombre modeste comparé aux près de 7 000 satellites Starlink en orbite. Cela dit, le rythme de remplacement des satellites de la constellation Starlink est rapide*

et ce nombre devrait croître très rapidement ». Mécaniquement, cette couverture va s'accroître avec le remplacement des satellites, mais Elon Musk ne va certainement pas se limiter à de simples échanges de SMS. Déjà, SpaceX a publié une démonstration d'un échange en visio depuis un smartphone standard. Si la FCC l'autorise à augmenter la puissance d'émission de ses satellites, Starlink pourrait bien offrir des services en concurrence directe avec les opérateurs… *« Les ambitions d'Elon Musk ne s'arrêteront sans doute pas au simple échange de SMS »*, estime Renaud Kayanakis. *« Il sait jouer sur la séduction en proposant des récepteurs Starlink aux autorités lorsque surviennent des catastrophes naturelles. La menace est d'autant plus sérieuse qu'il peut aujourd'hui s'adosser à l'exécutif américain. Les opérateurs ont de quoi s'inquiéter pour le futur. »*

Le business model des opérateurs bousculé ?

Tout l'enjeu pour les constructeurs de mobiles, les opérateurs cellulaires et les opérateurs de satellites va être de trouver un modèle économique pérenne, et Starlink arrive comme un chien dans un jeu de quilles. Les opérateurs misaient sur la dernière évolution de la 5G. La Release 17 de la spécification 3GPP intègre les NTN (Non-Terrestrial Network), les satellites géostationnaires pour compléter leurs infrastructures terrestres. Lors du MWC (Mobile World Congress) 2025, l'américain Skylo a justement démontré ses services de SOS et SMS en zone blanche. Une première démonstration avait été menée quelques mois plus tôt avec l'opérateur grec Cosmote, filiale de Deutsche Telekom. La communication a mis en œuvre une puce 5G Snapdragon X80, conforme à la spécification 3GPP Release 17, et implémentant la fonction Direct-to-Handset (D2H) du standard. Ce service est proposé par Verizon aux États-Unis, et la startup a signé un premier partenariat en Europe avec Deutsche Telekom.



Dans une vidéo publiée sur X, le 21 mai 2024, Starlink démontrait la faisabilité d'une visio réalisée en Direct-to-Cell sur un smartphone non modifié. La qualité n'est pas au rendez-vous, mais ça fonctionne !

Alors que 2025 sera la première véritable année « commerciale » du Direct-to-Cell, ce marché ne devrait générer que 30 millions de dollars au niveau mondial cette année. Le cabinet Juniper Research estime qu'il atteindra 1,7 milliard de dollars d'ici 2029. Ce chiffre est à mettre en regard aux 220 milliards de dollars qu'ont généré les réseaux cellulaires en 2024. C'est peu, très peu. Les opérateurs vont avoir beaucoup de mal à rentabiliser ces services facturés moins de 2 dollars par mois. Face à l'ogre Musk, qui cherche de nouveaux débouchés commerciaux pour sa constellation Starlink, le risque de disruption du marché est bien réel. □

A.C

« T-MOBILE FAIT ENTRER LE LOUP DANS LA BERGERIE ! »

RENAUD KAYANAKIS, ASSOCIATE PARTNER SPÉCIALISÉ DANS LES TÉLÉCOMS ET L'AUTOMOBILE CHEZ SIA PARTNERS

« Sur le papier, la proposition de valeur de Starlink est géniale : on dispose enfin d'une solution de communication pour les zones blanches. La solution semble fonctionner et il n'y a, a priori, aucun obstacle au déploiement d'un tel service qui n'est pas une alternative, mais un complément aux réseaux cellulaires. Tout le problème sera de trouver un équilibre concurrentiel entre les acteurs. En effet, les opérateurs mobiles ont acheté leurs licences plusieurs milliards de dollars. Ils sont soumis à des engagements, et ces règles doivent s'appliquer à tout le monde, or, l'arrivée de SpaceX dans la téléphonie va poser des problèmes d'équité et de droit. De même, les fréquences sont une denrée rare et les risques de brouillage des autres opérateurs sont réels. De même, la puissance d'émission pour échanger avec un satellite à plusieurs centaines de kilomètres en orbite va être bien supérieure à celle pour communiquer avec une antenne à quelques kilomètres. »

Routeur Netgear Orbi 970

Un kit mesh Wi-Fi 7 taillé pour l'avenir

Avec l'Orbi 970, Netgear passe un cap pour proposer un premier kit mesh haut de gamme compatible Wi-Fi 7. Optimisé pour les télétravailleurs et les PME, le système promet une couverture exceptionnelle, des débits fulgurants, et une gestion avancée du réseau. Un système taillé pour répondre aux besoins grandissants en termes de connectivité, mais qui se paie au prix fort.

Dès le premier coup d'œil, l'Orbi 970 impressionne par son format imposant. Chaque unité (un routeur et un ou deux satellites en fonction du kit choisi) arbore un châssis vertical robuste avec des dimensions bien plus volumineuses que celles des précédentes générations. Si ce gabarit peut compliquer l'intégration dans certains espaces de travail, il est justifié par un design thermique optimisé et la présence de douze antennes internes. La finition est impeccable, avec un revêtement (noir ou en blanc) sobre et premium semblant très robuste. Un support mural en option (49 €) est proposé pour faciliter son installation. L'un des atouts majeurs de ce kit est sa connectique particulièrement fournie, pensée pour les besoins des entreprises et des professionnels IT. Le routeur principal dispose de deux ports WAN/LAN de 10 Gbit/s, et de quatre ports Ethernet 2,5 Gbit/s permettant de relier des équipements critiques (serveurs, NAS, stations de travail) sans congestion. Le satellite embarque quant à lui un seul port WAN 10 Gbit/s, et deux ports Ethernet. De quoi assurer une interconnexion efficace entre différents postes sans dépendre uniquement du Wi-Fi. L'Orbi 970 se distingue de nombreuses solutions mesh du marché qui sont souvent limitées à une ou deux interfaces réseau.

Des performances Wi-Fi 7 hors normes

L'intégration du Wi-Fi 7 (802.11be) propulse les performances de l'Orbi 970 à un niveau inégalé. Relié à la box de l'opérateur, le routeur repose sur une configuration particulièrement musclée comprenant un processeur quadri cœur, 2 Go de mémoire vive et 4 Go de stockage dédiés au système d'exploitation. Grâce à une architecture quadri-bande (2,4GHz, 5GHz-1/5GHz-2 et 6GHz), l'ensemble est capable de délivrer des débits théoriques dépassant les 27 Gbit/s cumulés ! Netgear utilise la technologie MLO (Multi Link Operation) pour pouvoir utiliser simultanément la puissance des deux bandes Wi-Fi 5GHz et 6 GHz et obtenir des performances inédites. Le Wi-Fi 7 réduit considérablement la latence, un avantage clé pour les usages exigeant une réactivité sans faille, comme la visioconférence en haute définition, la production 3D, ou encore les applications de réalité virtuelle et augmentée. Seules les meilleures box fibre (Livebox Max Fibre, Freebox Ultra…) permettent actuellement d'exploiter pleinement les performances de l'Orbi 970. De plus, pour atteindre les débits optimaux offerts par le Wi-Fi 7, il est indispensable de disposer d'appareils compatibles avec cette norme, sans quoi les vitesses resteront limitées aux capacités des générations précédentes. Lors de nos tests avec une Freebox Ultra et un iPhone 16 Pro, nous avons mesuré des débits allant jusqu'à 2,3 Gbit/s en téléversement et téléchargement.

Un backhaul Wi-Fi 7 dédié

Afin d'assurer une interconnexion fluide, Netgear a intégré un backhaul 10 Gbit/s dédié en Wi-Fi 7 : un canal spécifiquement réservé à la communication entre le routeur et

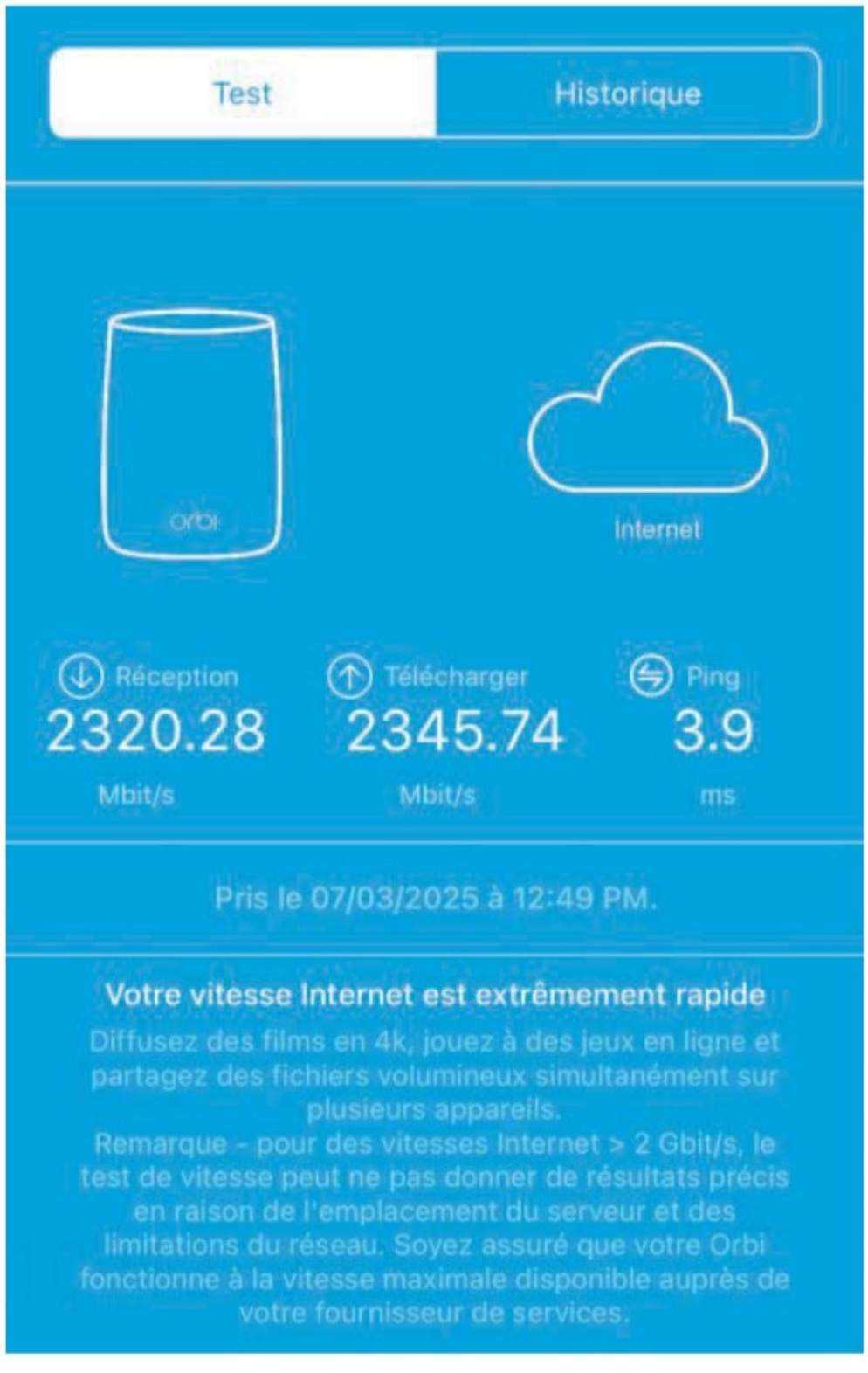
le satellite. Cette technique permet d'éviter que les performances ne soient affectées par la charge du réseau principal, tout en garantissant une latence minimale et un débit stable sur l'ensemble de la couverture. Les ports 10 Gbit/s assurent un backhaul filaire robuste et fiable avec une bande passante maximale. Une connectique optimisée pour relier des équipements critiques (serveurs, NAS, stations de travail) sans congestion. Avec le kit comprenant un routeur et un satellite, la couverture peut atteindre jusqu'à 460 m², tandis qu'en optant pour la version avec deux satellites, elle s'étend jusqu'à environ 660 m². L'ajout d'un satellite autonome (900 €) étend encore la portée de 220 m² supplémentaire. Quel que soit le nombre de modules, il est possible de connecter jusqu'à 200 appareils au total.

Une couche de cybersécurité renforcée

Netgear accompagne son offre d'un an d'abonnement gratuit à Netgear Armor, une suite de cybersécurité développée en partenariat avec Bitdefender. Cette solution protège l'ensemble des appareils connectés au réseau contre les menaces, incluant les malwares, les ransomwares et le phishing. L'intégration de ce service au niveau du routeur permet de sécuriser les équipements sans nécessiter d'installation logicielle supplémentaire sur chaque machine. Il est toutefois possible de sécuriser jusqu'à 50 périphériques en déplacement (smartphones, tablettes, PC portables, MacBook...) en installant cette fois-ci l'application Netgear Armor. Un atout non négligeable pour les entreprises soucieuses de renforcer la protection de leurs infrastructures face aux cyberattaques.

Une application de gestion avancée

L'application mobile Orbi jouit d'une interface intuitive pour gérer le réseau en quelques clics. Elle permet de configurer plusieurs SSID pour faciliter la segmentation entre un réseau principal, un réseau invité et un espace dédié aux appareils IoT. Cette séparation est essentielle pour limiter les risques de compromission des objets connectés souvent moins sécurisés. Grâce à ses diagnostics en temps réel



Les débits sans fil offerts par l'Orbi 970 avec des périphériques compatibles tiennent toutes leurs promesses.

et ses alertes de sécurité, l'application facilite la gestion du réseau, y compris pour les non-experts IT. L'installation du kit est également simplifiée grâce à une amélioration du logiciel de configuration, qui intègre un système de QR codes et un guide pas à pas très intuitif. La mise en route est relativement simple et rapide, mais il faut prévoir un certain laps de temps avant que le réseau ne se stabilise parfaitement.

CARACTÉRISTIQUES TECHNIQUES
NETGEAR ORBI 970 (KIT DE 2)

- **Processeur :** Quad-Core 2,2 GHz
- **Mémoire vive (RAM) :** 2 Go
- **Stockage :** 4 Go
- **Nombre d'antennes :** 12
- **Bandes de fréquence :** 2,4 GHz, 5 GHZ, et 6 GHz
- **Débits théoriques :** 27 GHz cumulés
- **Connectique routeur principal :** 1 x 10 Gbit/s WAN, 1 X 10 Gbit/s LAN, 4 x 2,5 Gbit/s Ethernet
- **Connectique satellite :** 1 x 10 Gbit/s WAN, 2 x 2,5 Gbit/s Ethernet
- **Dimensions (l x l x h) :** 29,4 x 14,44 x 13,1 cm
- **Poids :** 1,79 kg
- **Tarif lancement :** 1699 €

Conclusion

Le Netgear Orbi RBE970B s'impose comme l'une des solutions Wi-Fi 7 les plus performantes du marché. Sa puissance et sa connectique en font un choix particulièrement adapté aux PME, aux bureaux partagés et aux environnements complexes nécessitant une connectivité fiable et rapide. Son prix élevé et son encombrement imposant peuvent toutefois constituer un frein. L'absence d'un port USB, qui aurait facilité la connexion d'un disque dur ou d'un NAS, est également regrettable. Sans un réseau fibre haut de gamme et un parc d'appareils compatibles Wi-Fi 7, il sera difficile d'exploiter pleinement son potentiel. L'Orbi 970 constitue donc un investissement d'avenir, à condition d'avoir une infrastructure adaptée pour en tirer le meilleur parti. □

J.C

Qualtrics X4 2025

Redéfinir l'expérience client et employé avec l'IA

Qualtrics intègre de plus en plus l'IA pour améliorer l'expérience client et employé. Lors de Qualtrics X4, l'entreprise a présenté ses Experience Agents, des agents autonomes qui détectent et résolvent les problèmes en temps réel. Elle a aussi lancé Qualtrics Edge, une plateforme d'intelligence de marché exploitant l'IA pour faciliter la prise de décision. Enfin, en partenariat avec LangChain, Qualtrics entend optimiser l'intégration de ses solutions dans les systèmes existants.

Aujourd'hui, la rapidité et la personnalisation sont devenues des facteurs décisifs dans la fidélisation des clients et l'engagement des employés. Afin de répondre à ces attentes, l'éditeur a intégré l'intelligence artificielle au cœur de son approche. Lors de Qualtrics X4 2025 (18-20 mars à Salt Lake City), l'entreprise a dévoilé les Experience Agents, des agents autonomes capables de détecter, comprendre et résoudre les problèmes en temps réel à chaque point de contact avec une entreprise.

Cette avancée s'inscrit dans une transformation plus large, portée par le lancement de Qualtrics Edge, une nouvelle plateforme d'intelligence de marché qui promet d'accélérer la prise de décision et d'améliorer la compétitivité des entreprises. Avec ces innovations, Qualtrics veut bouleverser les paradigmes traditionnels en remplaçant les approches réactives par une gestion proactive et prédictive des expériences. En s'appuyant sur des technologies avancées d'intelligence artificielle et d'analyse de données, l'entreprise permet aux organisations de comprendre leurs clients et employés, d'anticiper leurs besoins et d'agir au moment où cela compte le plus.

Des agents autonomes proactifs

La plupart des entreprises s'appuient sur des enquêtes et des analyses post-événement pour comprendre les attentes de leurs clients. Ce modèle présente un inconvénient majeur : les problèmes ne sont détectés qu'après qu'ils ont eu un impact négatif, ce qui réduit les chances de rétention et d'amélioration rapide. Avec ses Experience Agents, l'éditeur change la donne en permettant aux entreprises de réagir instantanément aux frustrations et attentes des consommateurs.

« Ces agents sont une extension de votre force de travail », a expliqué Zig Serafin, PDG de Qualtrics. *« Ils apportent un niveau inédit d'intelligence, de prise de décision et d'action. Contrairement aux agents transactionnels classiques ou aux chatbots basés sur des règles, ils s'appuient sur la capacité unique à comprendre les personnes. »* Plutôt que de collecter simplement des réponses dans un questionnaire ou un avis en ligne, ces agents autonomes analysent le langage naturel, détectent les émotions et interprètent le contexte afin de fournir une réponse personnalisée et contextuelle. Si un client mentionne un problème dans un sondage, l'agent ne se contente pas de noter l'insatisfaction, il intervient immédiatement en proposant une solution adaptée : un remboursement, une réduction, un suivi prioritaire ou toute autre action corrective qui peut transformer une expérience négative en opportunité de fidélisation.

L'un des aspects de ces agents est leur capacité à fonctionner sur plusieurs canaux simultanément. Ils peuvent intervenir non seulement sur les enquêtes de satisfaction, mais aussi sur les avis clients en ligne, les interactions sur les réseaux sociaux ou les expériences sur un site web. Cette approche omnicanale permet aux entreprises de fermer la boucle plus rapidement, réduisant ainsi le taux d'attrition et améliorant la perception de leur marque. Par ailleurs, Qualtrics va encore plus loin en intégrant ces capacités dans une vision prédictive. En analysant l'historique des interactions et les tendances de comportement, les Experience Agents pourront anticiper les attentes et intervenir avant même qu'un client n'exprime un mécontentement. Cette évolution représente une avancée majeure, en transformant les expériences passives en expériences proactives et engageantes.

© Qualtrics

Zig Serafin, PDG de Qualtrics, lors du keynote d'ouverture de Qualtrics X4 à Salt Lake City (18-20 mars)

Une évolution de l'intelligence de marché

En parallèle des Experience Agents, Qualtrics introduit Qualtrics Edge, une plateforme qui permet aux entreprises d'accéder instantanément à des informations stratégiques, tout en réduisant les coûts et les délais des études de marché. Traditionnellement, la collecte d'informations sur les tendances de marché, la satisfaction client et l'analyse concurrentielle repose sur des études longues et coûteuses, souvent limitées par la taille et la représentativité des échantillons. Qualtrics Edge fait la différence en combinant des données synthétiques, des benchmarks compétitifs et des capacités avancées d'IA pour offrir une visibilité instantanée sur l'évolution du marché. L'une des innovations majeures de cette plateforme réside dans l'introduction de Edge Audiences, un outil qui remplace les panels traditionnels par des audiences synthétiques générées par IA. Cette technologie permet aux chercheurs d'accéder en quelques minutes à des informations représentatives de leur marché cible, sans passer par des processus longs et onéreux. Les modèles d'apprentissage automatique développés par Qualtrics sont capables de simuler des comportements et préférences avec une grande précision, permettant aux entreprises d'obtenir des réponses exploitables immédiatement.

Par ailleurs, Qualtrics Edge Instant Insights fournit un accès en temps réel à des benchmarks sectoriels et à des tendances de consommation, offrant aux entreprises la possibilité de comparer leurs performances avec celles de leurs concurrents et d'identifier rapidement les opportunités de croissance. Cette fonctionnalité, alimentée par des millions de points de données collectés à travers diverses industries, permet aux entreprises de prendre des décisions stratégiques. L'avantage clé de Qualtrics Edge réside dans sa capacité à transformer des données complexes en recommandations actionnables, grâce à l'intelligence artificielle. En croisant les informations provenant des clients, des concurrents et des tendances du marché, l'outil identifie les actions prioritaires à mettre en place, réduisant ainsi le risque d'erreur dans la prise de décision et accélérant l'adaptation aux évolutions du marché. Concernant ces technologies, les premiers utilisateurs peuvent déjà en profiter, et la disponibilité générale est prévue d'ici la fin de l'année en français et d'autres langues.

Partenariat stratégique avec LangChain

Pour maximiser l'efficacité de ses Experience Agents, Qualtrics annonce un partenariat avec LangChain, un spécialiste des modèles de langage avancés (LLM). Grâce à la plateforme LangGraph, Qualtrics peut déployer et gérer ses agents autonomes de manière plus fluide et efficace. L'un des principaux défis des entreprises adoptant l'IA conversationnelle réside dans l'intégration de ces agents avec les autres technologies existantes. En collaborant avec LangChain, Qualtrics s'assure que ses agents peuvent fonctionner de manière transparente avec les infrastructures technologiques déjà en place, maximisant ainsi leur impact et leur interopérabilité.

© MC

Brad Anderson, président des produits et de l'ingéniérie de Qualtrics, est intervenu le premier jour de Qualtrics X4 pour détailler les principales nouveautés de l'entreprise comme les Experience Agents, Qualtrics Edge ou encore les nouvelles fonctionnalités de XM for Employee Experience.

« Les Experience Agents représentent un changement dans la manière dont les entreprises interagissent avec leurs clients et employés, rendant possible une amélioration continue de leur expérience sur tous les canaux et points de contact », a déclaré Gurdeep Singh Pall, président de la stratégie IA de Qualtrics. *« Le partenariat avec des organisations de premier plan comme LangChain crée un écosystème ouvert pour le développement des agents »*, a-t-il souligné.

Une transformation de l'expérience employé grâce à l'IA

Qualtrics applique ces mêmes principes d'intelligence artificielle à l'expérience employé, un domaine où la compréhension et l'engagement des collaborateurs sont essentiels à la productivité et à la rétention des talents. Les nouvelles fonctionnalités de XM for Employee Experience permettent aux entreprises de capturer, analyser et exploiter les retours des employés pour améliorer leur satisfaction et leur engagement. Plutôt que de s'appuyer uniquement sur des enquêtes annuelles, Qualtrics permet aux organisations d'avoir un suivi continu et dynamique du moral et des attentes des employés. Grâce à des outils comme Qualtrics Assist, l'IA peut analyser des milliers de retours écrits, détecter les tendances et formuler des recommandations personnalisées aux managers. Cette approche permet de mieux comprendre les sources de motivation et d'insatisfaction, et d'agir en conséquence pour éviter les démissions et améliorer la cohésion des équipes.

Avec ses innovations en intelligence artificielle, Qualtrics entend mettre en avant un nouveau standard dans la gestion de l'expérience des clients et des employés. Dans un monde en constante évolution, Qualtrics cherche ainsi à montrer que l'avenir de l'expérience client et employé repose sur l'intelligence artificielle, l'anticipation et la personnalisation à grande échelle. □

M.C

Cloud

Oracle avance ses arguments pour l'IA en France

L'éditeur américain mise sur OCI, ses bases de données et sa capacité à s'adapter aux besoins spécifiques des entreprises françaises pour devenir un partenaire clé dans l'IA. Des clients comme BNP Paribas et Alstom se sont déjà engagés.

L'IA prend le relais du cloud pour offrir une nouvelle jeunesse à Oracle sur le marché français. Si l'éditeur américain avait eu du mal à prendre le tournant du cloud au début, il a su vite rebondir, comme en témoignent encore les résultats de son premier trimestre 2025, qui a vu les activités IaaS et SaaS progresser de 22 % et OCI (pour Oracle Cloud Infrastructure) de 46 %. Et il est hors de question pour l'éditeur d'avoir le même retard à l'allumage avec l'IA. C'est en tout cas l'ambition affichée par Christophe Négrier, senior vice-president Tech Cloud pour l'Europe du sud et Country Leader d'Oracle France lors de l'édition du Oracle World Tour, qui s'est tenue à Paris début mars.

Christophe Negrier, Oracle France country leader & SVP Tech EMEA South

« En France, l'intelligence artificielle générative devrait générer entre 250 et 420 milliards d'euros de PIB en plus d'ici 8 ans. Or aujourd'hui, seules 10 % des entreprises françaises ont réellement adopté cette dernière en production », rappelle le dirigeant. *« Nous avons une volonté farouche d'être aux côtés des entreprises et des administrations pour les aider à tirer parti de cette révolution. »*

OCI est taillé pour l'IA

Tout cela pourrait rester de belles paroles, mais Oracle a déjà démontré sa capacité à fournir des infrastructures capables de répondre aux besoins de l'IA, et certains de ses clients français sont déjà bien avancés dans leurs projets sur le sujet. Côté infrastructure, Oracle avance conjointement son offre OCI et ses produits de base de données, notamment Oracle Database (relationnelle) et Autonomous Database (Data Warehouse, Data Lake, etc.), qu'il est maintenant possible d'héberger à peu près partout. *« OCI, et plus particulièrement l'itération Gen2 Cloud, a été construite sur les bases réseau de nos baies hautes performances Exadata, associées à des technologies Nvidia, en faisant de fait une infrastructure taillée pour l'IA »*, rappelle Sudha Raghavan, SVP Developer Platform d'Oracle. Et pour cause, c'est sur cette infrastructure qu'ont été développés et entraînés des LLM comme, entre autres, ChatGPT et Llama 3.

Au-delà de son infrastructure, le grand avantage d'Oracle réside aussi dans sa capacité à pouvoir déployer ses technologies dans de très nombreux modes (cloud privé, hybride, public…), avec une localisation précise des données. C'est ainsi que BNP Paribas a déployé les solutions de base de données Oracle en mode cloud, mais au sein de ses propres datacenters. La banque assure d'ailleurs aujourd'hui avoir plus de 800 cas d'usage d'IA en production, dont la plupart tournent sur Oracle. *« Nous allons construire, pour l'ensemble des métiers de la banque, une offre IA qui va ressembler à celle d'un cloud provider avec divers partenaires accrédités et vérifiés »*, raconte ainsi Jean-Michel Garcia, Group CTO de BNP Paribas.

Une réponse spécifique pour les problèmes de localisation des données

Du côté d'Alstom, c'est la flexibilité qui est avancée. L'industriel, présent dans 70 pays, a fait le choix d'héberger ses bases de données Oracle en interne, chez des hyperscalers (en l'occurrence Azure) ou sur des infrastructures OCI propres à Oracle. *« Nous avons eu le cas avec un client saoudien qui ne voulait pas que ses données sortent du pays. Nous avons ainsi localisé la stack de données sur des infrastructures Oracle dans le pays, tandis que le reste des solutions était hébergé ailleurs »*, explique Stéphane Detruiseux, VP Information Security et Infrastructure d'Alstom. Pour appuyer la pertinence de ses offres en matière de localisation des données, Oracle a d'ailleurs tenu à rappeler qu'il avait abaissé à « seulement » trois racks l'emprise minimum pour pouvoir déployer sa région dédiée d'OCI chez soi.

Avec ses arguments, Oracle espère ainsi s'imposer comme un partenaire idéal pour l'IA en France, d'autant plus que la firme a déjà le pied dans la porte chez de nombreux comptes. *« Plus de 90 % des entreprises du CAC 40, des administrations et des ministères nous font déjà confiance »*, rappelle Christophe Négrier. □

O. Ba

Communauté

Agentforce ouvre sa boutique

Alors que s'est tenue la conférence développeur de Salesforce à San Francisco, l'éditeur annonce l'ouverture de sa boutique d'agent sur le modèle d'AppExchange.

Il sera loisible aux partenaires et aux clients de mettre à dispositions différents types de composants d'agents IA sur la boutique et de les monétiser s'ils le souhaitent. Ils peuvent ainsi proposer des agents prenant des actions qui étendent les capacités des agents en ajoutant de nouvelles intégrations — allant d'Apex, de flows, d'API et de prompts — pour adapter des composants spécifiques à chaque secteur. Des modèles de prompts offrant des instructions réutilisables et optimisées pour garantir des interactions cohérentes peuvent être mis à disposition des utilisateurs. Les agents seront réunis par rubriques pour organiser et affiner le comportement des agents en regroupant les actions et les instructions autour d'une tâche ou d'une mission. Cela permet de garantir que les agents produisent des résultats cohérents et respectent des directives prédéfinies. Des modèles d'agents vont eux fournir des solutions d'IA complètes en combinant plusieurs rubriques et en utilisant les actions créées par les partenaires, accompagnées de métadonnées et d'instructions globales couvrant l'ensemble des rubriques. Les clients peuvent explorer les solutions AgentExchange directement sur la marketplace ou via l'outil Agent Builder de Salesforce pour identifier les solutions adaptées à leur cas d'usage, produit ou secteur. 200 de ces composants sont déjà présents sur la boutique.

Patrick Stokes, EVP Products chez Salesforce, lors de sa session plénière

Ils répondent à des besoins selon des secteurs d'activité ou des fonctions comme la signature électronique.

Les clients ont le choix de passer directement par l'éditeur du composant ou de passer par la boutique. Au moment du choix, un pop-up propose l'installation directement à partir de la marketplace, ce qui donne accès à un administrateur pour finir l'installation. Celui-ci reçoit une notification de la demande de l'utilisateur pour ce composant et finalement l'installe. Au passage, Salesforce prend 15 points sur le prix du composant, le reste étant payé par Salesforce à l'éditeur du composant. C'est le même modèle que sur AppExchange. Selon Nick Johnston, en charge des partenariats stratégiques chez Salesforce, ce modèle est appelé à évoluer selon les retours des partenaires, les composants téléchargés. In fine, le modèle devrait se rapprocher de ce que proposent les hyperscalers dans ce domaine. La boutique est d'ores et déjà en service. Le packaging et le référencement des modèles de prompts et des rubriques sont disponibles dès aujourd'hui. Le packaging et le référencement des modèles d'agents seront disponibles en avril 2025.

À l'analyse, il semble que cette ouverture sur les agents ouvre la voie à une sorte de standardisation des composants d'agents IA, afin de faciliter leur collaboration. Nous devrions avoir plus d'éléments en ce sens lors du prochain Dreamforce à l'automne prochain. Nick Johnston entrevoit cette possibilité de développer une sorte de manière commune de faire travailler ensemble les agents ou les API. □

B.G

SALESFORCE ÉTEND SON PARTENARIAT AVEC GOOGLE CLOUD

Lors de la conférence développeur TDX 25, Salesforce a annoncé le renforcement de son partenariat avec Google. Les deux sociétés travaillent ensemble sur différents sujets depuis près de 7 ans. Le partenariat actuel se limitait le plus souvent à une bonne intégration et à ce que les produits Salesforce et Google travaillent bien ensemble. Selon Nick Johnston, en charge des partenariats technologiques stratégiques de Salesforce, l'annonce va beaucoup plus loin et se propose de porter l'ensemble des produits cœur de l'éditeur sur le cloud de Google. Actuellement, les produits sont sur AWS. Nick Johnston indique : *« cela vient de la demande récurrente de clients qui souhaitaient utiliser nos logiciels sur l'hyperscaler que j'ai choisi. Cela nous permet de proposer plus de choix à nos clients, mais aussi de gérer des problématiques de résidence des données. Nous aurons la possibilité d'aller dans des lieux où nous n'avons pas de centre de données »*. La solution est en cours de réalisation et sera disponible l'année prochaine.

Agentique

AWS se met en ordre de bataille

Contrairement à ses rivaux Google et Microsoft, Amazon et AWS n'ont pas choisi de se lancer dans la grande bataille des LLM. L'engouement suscité par ChatGPT, et plus largement par l'IA générative, a plongé l'hyperscaler dans l'ombre, mais il compte bien se refaire sur la nouvelle révolution à venir, l'IA agentique.

Face à la déferlante de LLM initiée par OpenAI en 2022 avec le succès météorique de ChatGPT, AWS a fait partie des acteurs du Cloud à ne pas s'être lancé immédiatement dans la bataille. Comme Oracle ou, en France, OVHcloud, AWS a adopté une posture de relative neutralité, en proposant à ses clients d'héberger les LLM de leur choix. Une stratégie prudente lorsqu'on connaît le coût d'apprentissage de nouveaux modèles fondation. Microsoft a ainsi injecté plus de 10 milliards dans OpenAI, et Google a surenchéri au début de l'année en annonçant 75 milliards d'investissement dans l'IA générative.

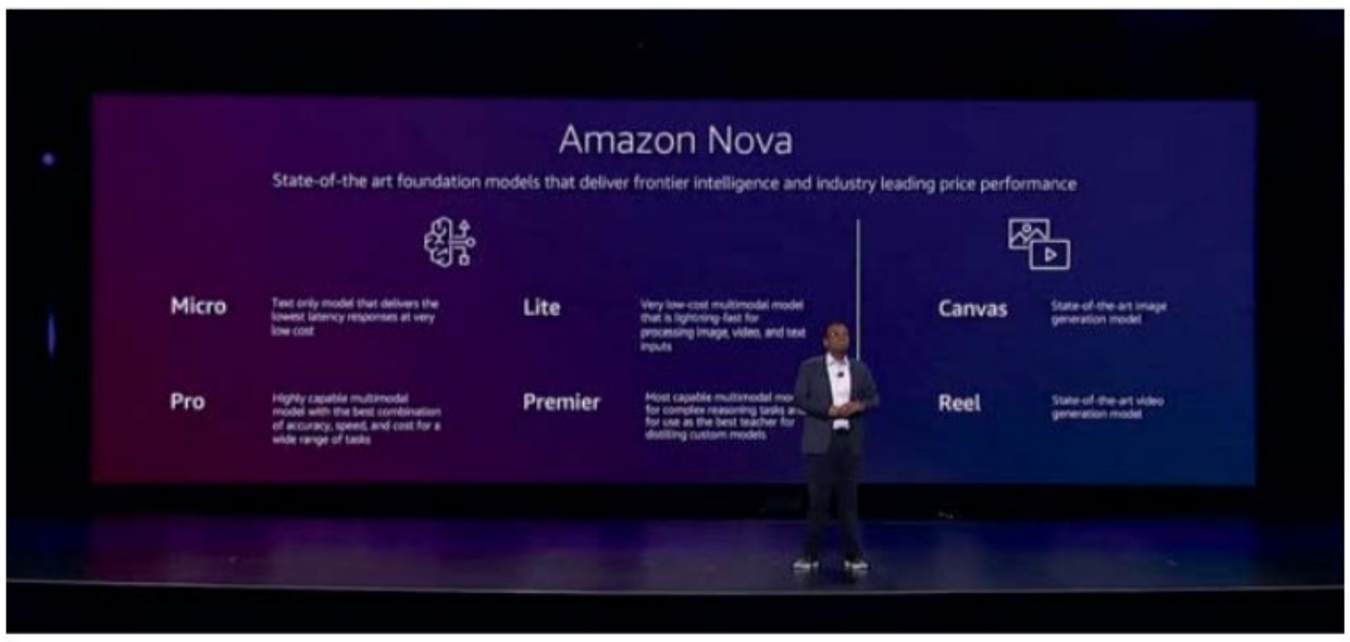
Andy Jassy, CEO d'Amazon, a annoncé lors de la présentation des résultats d'Amazon au quatrième trimestre, un investissement de 100 milliards de dollars dans l'IA sur l'année. Une surenchère en réponse à l'annonce des 500 milliards du plan Stargate de Donald Trump.

Ces dernières années, AWS s'est construit un catalogue de solutions IA plutôt solide. SageMaker et sa récente évolution Hyperpod couvrent tout le cycle de vie des modèles analytiques et des IA génératives, Bedrock permet de construire des applications d'IA générative, notamment les fameux RAG, avec un accès aisé aux LLM créés par AI21labs, Anthropic, Cohere, Mera, Mistral AI, Poolside, et Stability ai. La marketplace Amazon Bedrock compte aujourd'hui plus de 100 LLM. Si on ajoute à ces solutions majeures le moteur de recherche intelligent Amazon Kendra ou encore Amazon Q, son assistant intelligent, AWS a assemblé un beau catalogue de solutions IA. On peut ajouter l'effort réalisé dans le hardware pour s'affranchir, en partie, de la domination de Nvidia, avec la mise au point des composants spécialisés Inferencia et Trainium3. Au dernier trimestre 2024, AWS a injecté 26,3 milliards de dollars dans son infrastructure cloud pour capter ce marché de l'IA, et compte maintenir ses investissements au rythme de 100 milliards par an.

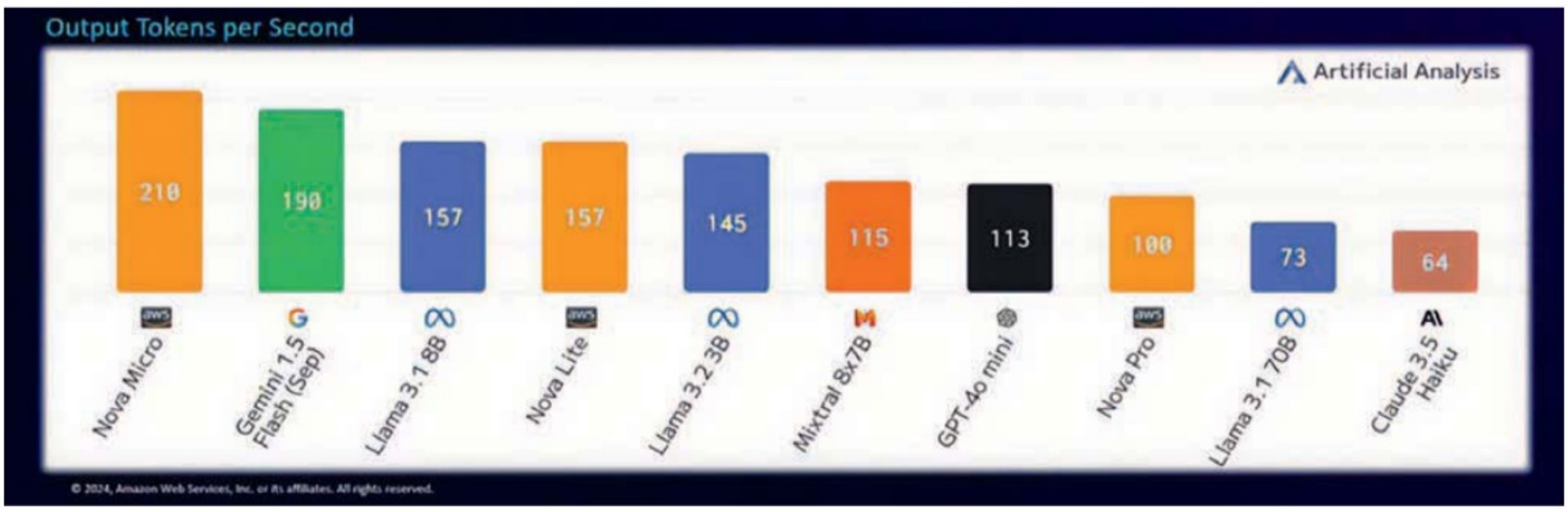
AWS est devenu acteur

Fournir des infrastructures, c'est bien, mais rester en marge de la grande bataille des LLM présentait le risque de passer pour un acteur de second plan aux yeux du Nasdaq. Devant la surenchère des milliards initiée par Microsoft, Google, puis le plan Stargate de 500 milliards de dollars annoncé par Donald Trump avec Softbank, OpenAI et Oracle en embuscade, AWS courait ce risque précisément.

Pour l'heure, c'est Microsoft qui profite à plein des synergies entre OpenAI et Azure, toutes les grandes entreprises qui ont souhaité déployer un ChatGPT interne l'ont fait sur leur tenant Azure pour que leurs données ne fuitent pas sur le service grand public.



Lors de la conférence annuelle AWS re : Invent en 2024, Swami Sivasubramanian, encore responsable data et IA, présentait la gamme de LLM Nova. Il est aujourd'hui aux commandes de la stratégie IA agentique de l'hyperscaler.



AWS présente Nova Micro comme un LLM a faible latence pour traiter du texte rapidement. C'est Nova Pro et, a fortiori, le futur Nova Premier, qui seront les modèles fondation les plus adaptés aux IA agentiques autonomes.

En décembre dernier, lors de son grand événement annuel re: Invent, AWS annonçait ses modèles fondation Amazon Nova. Plutôt que de chercher à rivaliser en milliards de paramètres d'apprentissage, AWS a préféré livrer son LLM ou plutôt ses LLM en trois éditions : Micro pour le texte uniquement, Lite, une IA générative multimodale pour le texte, l'image et les vidéos et enfin Pro, un LLM plus polyvalent. Ces modèles seront ensuite complétés par l'édition Premier, le haut de gamme dont on ne sait rien de plus. Ces modèles « maison » sont accessibles via Amazon Bedrock, aux côtés des LLM édités par des tiers.

Si Amazon a quelque peu raté le coche sur les chatbots d'entreprise et les RAG, avec Nova, l'américain vise le coup d'après, l'IA agentique. Ses IA génératives doivent notamment motoriser les agents intelligents, la fameuse IA agentique qui s'annonce comme le buzzword de ces prochains mois. L'incontournable cabinet Gartner annonce qu'en 2028, 33 % des applications d'entreprises embarqueront de l'IA agentique dans leurs workflows. Les cadres seront aussi directement concernés par cette automatisation puisque 15 % des décisions prises au jour le jour le seront par des agents de manière autonome. En front office, ces agents remplaceront de l'ordre de 20 % des interactions.

Si ces prédictions se matérialisent vraiment, l'IA agentique va avoir un impact considérable sur les entreprises. AWS, qui compte 80 % du CAC 40 parmi ses clients, espère bien que ces agents autonomes seront créés sur son architecture. L'IA agentique lui donne une belle occasion de rebondir et la gamme Nova arrive à point nommé pour motoriser son offre agentique.

Amazon Bedrock Agents permet de créer des agents « intelligents » capables de réaliser des traitements relativement complexes avec plusieurs étapes et mettant en jeu plusieurs sources de données. Avec Nova et son édition Premier, capable de raisonnement complexe, il serait possible de créer des agents autonomes.

Parmi les premiers partenaires présentés par AWS autour de ses modèles fondation, figure SAP. L'éditeur compte intégrer les LLM Nova à son Generative AI Hub. De même, Palantir compte intégrer Nova dans la brique Ontology de sa plateforme. Enfin, Deloitte a annoncé son intention de s'appuyer sur les LLM Nova pour créer les applications customisées pour ses clients.

Une nouvelle direction dédiée à l'IA agentique

Outre cette nouvelle stratégie produit, Reuters a révélé qu'AWS en a créé une nouvelle dédiée à l'IA Agentique, dont l'objectif sera d'aider les entreprises à créer leurs agents. A la différence d'Oracle et de Microsoft qui sont à la fois fournisseurs de services Cloud, mais aussi éditeurs d'applications SaaS, AWS n'a pas de compétences internes sur les process d'entreprise. Ces compétences métiers sont pour partie chez les intégrateurs, chez les éditeurs de solutions métiers, et chez ses clients. AWS doit donc trouver le moyen de se rendre indispensables aux entreprises qui voudront se doter d'agents autonomes.

Dans un email révélé par Bloomberg, Matt Garman, CEO d'AWS pointe l'importance des agents dans la stratégie d'AWS : *« L'IA agentique a le potentiel d'être la prochaine activité multimilliardaire d'AWS [...] Nous avons la possibilité d'aider nos clients à innover encore plus rapidement et à exploiter davantage de possibilités, et je suis fermement convaincu que les agents d'IA sont au cœur de cette prochaine vague d'innovation. »* La conséquence de ce repositionnement stratégique serait un vaste remaniement des équipes en charge des produits IA d'AWS. Le site américain évoque ainsi un message interne envoyé par Peter DeSantis, senior vice-president d'AWS qui annonce un regroupement des équipes Bedrock, SageMaker, mais aussi des ingénieurs en charge du Hardware, afin de créer une Task Force dédiée à l'IA agentique. Swami Sivasubramanian, auparavant vice-président Data et IA d'AWS vient d'être nommé VP Agentic AI, afin de mener cette offensive sur un marché qui représente de l'ordre de 5 milliards de dollars en 2024, mais qui pourrait offrir un taux de croissance de 30 à plus de 40 % par an jusqu'en 2030, selon les analystes. Une belle bataille en perspective entre les fournisseurs Cloud qui ont la puissance de traitement, mais pas de compétences métiers, et les éditeurs de logiciels qui sont dans une situation inverse. □

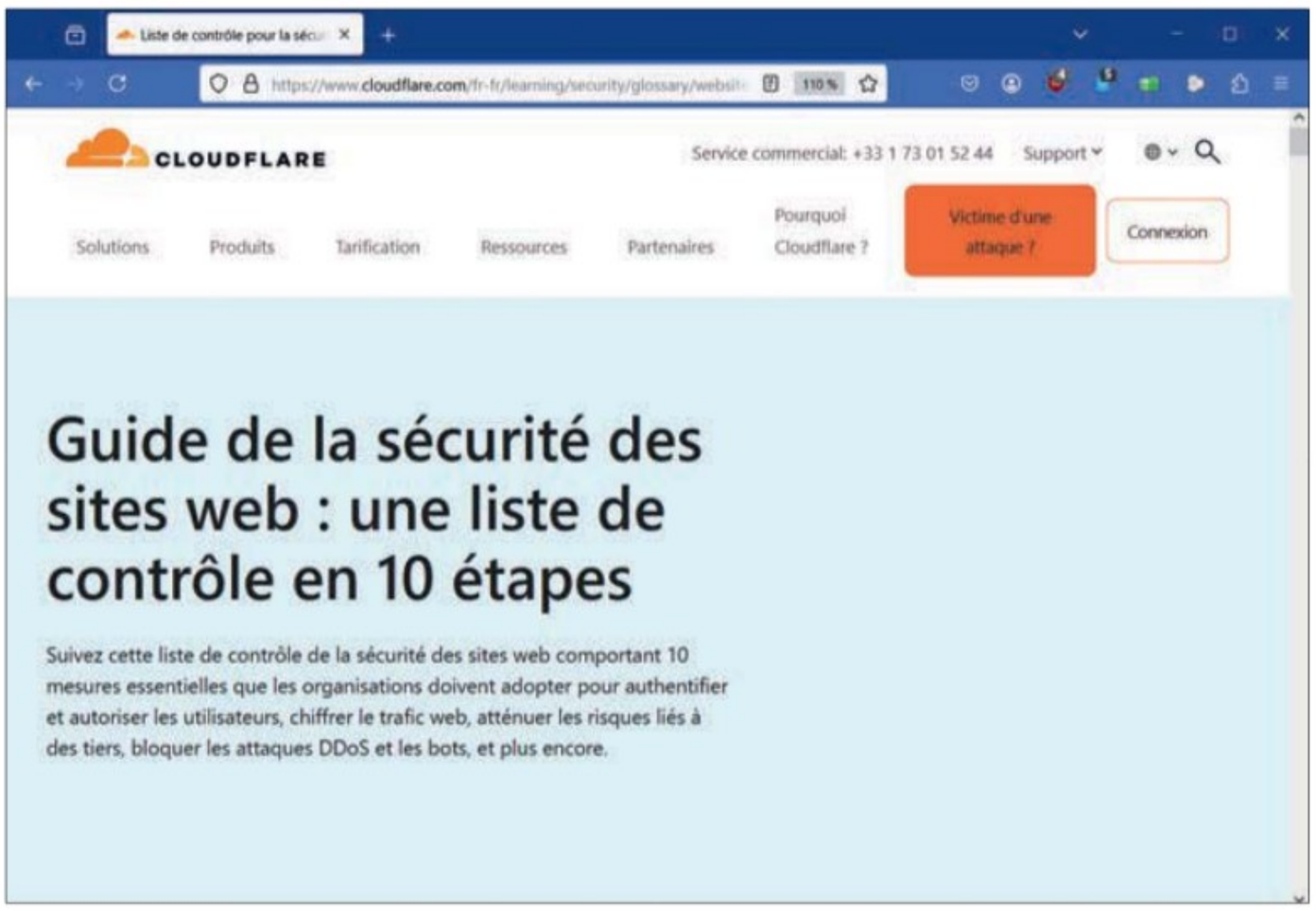
A.C

Applications

Cloudflare muscle son WAF pour protéger les LLM

Cloudflare a ajouté une fonction de sécurisation des LLM dans son WAF (Web Application Firewall). Appelée Firewall for AI, elle analyse les requêtes et évite leur exploitation malveillante. Quels sont les principes de son fonctionnement ?

Les spécialistes du cloud et de la cybersécurité se penchent, telles de bonnes fées, sur la protection des modèles de langage de grande taille, les fameux LLM (large language model). Le dernier en date est Cloudflare qui a mis à jour son WAF — son firewall applicatif — en y ajoutant une fonction baptisée Firewall for AI. Celle-ci a été spécialement conçue pour les applications utilisant des LLM. Elle est constituée d'outils WAF existants et nouveaux, capables d'analyser les invites soumises et d'identifier les tentatives d'exploitation frauduleuses. Avec Firewall for AI, Cloudflare propose la protection des LLM aux utilisateurs de Workers AI, sa plateforme serverless pour développeurs. Cette barrière de sécurité a pour fonction de détecter les tentatives d'attaques avant qu'elles ne ciblent ces modèles d'IA spécialisés dans l'interprétation du langage humain et d'autres formes de données complexes. Le WAF met en œuvre les informations sur les menaces et l'apprentissage automatique (machine learning), le tout soutenu par l'intelligence de plateforme du cloud de connectivité de Cloudflare, pour bloquer les dernières menaces, y compris zero-day. Le réseau mondial du fournisseur de services traite quelque 100 millions de requêtes HTTP par seconde (en pic). Cela lui permet clairement de proposer une protection inégalée contre les attaques, même les fameuses zero-day. *« À chaque avancée technologique correspondent de nouvelles menaces. Ce constat est également valable pour les technologies propulsées par l'IA. Avec notre service Firewall for IA, nous entendons intégrer la sécurité dès le début dans l'écosystème de l'IA »*, a déclaré Matthew Prince, cofondateur et PDG de Cloudflare. Du fait de leur capacité à traiter d'énormes quantités de données et à générer du texte, les LLM sont exposés à plusieurs types d'attaques bien spécifiques. Ces attaques peuvent viser soit à exploiter les vulnérabilités inhérentes à ces modèles, soit à profiter de leur fonctionnement pour mener à bien des actions malveillantes. Les attaques par injection de données, par exemple, consistent à introduire des données malveillantes ou trompeuses durant le processus d'apprentissage du modèle, dans le but de le manipuler pour qu'il génère des réponses biaisées ou inappropriées. Les modèles peuvent également être exploités pour générer du contenu trompeur comme des fake news, des escroqueries par hameçonnage ou toute autre forme de contenus malveillants en exploitant leur capacité à créer des éléments crédibles, texte, son ou vidéo. *« Firewall for AI est agnostique par rapport au déploiement spécifique. Il peut être employé pour protéger des modèles hébergés sur Cloudflare Workers AI, mais aussi sur n'importe quelle autre infrastructure tierce, du moment que le trafic transite en proxy via Cloudflare WAF »*, a encore déclaré le PDG du fournisseur de cloud. Firewall for AI fournit aux équipes de sécurité les outils nécessaires, afin de sécuriser leurs applications basées sur les LLM. En se positionnant en amont de n'importe quel LLM déployé sur la plateforme Workers AI, le pare-feu permet l'identification des menaces. Il a la capacité de repérer les tentatives malintentionnées d'exploitation des modèles, grâce à l'analyse et à l'évaluation des requêtes des utilisateurs.



Vous trouverez sur le site de Cloudflare un très bon guide de la sécurité des sites web avec une liste de contrôle en 10 étapes.

Protection contre le déni de service et les fuites de données

Firewall for AI est donc spécifiquement destinée aux clients qui exécutent une IA sur Workers AI. Elle protège notamment contre la fuite de données et l'injection d'invites. L'IA défensive analyse

et évalue les invites soumises par un utilisateur, afin de bloquer l'exploitation de modèles et les tentatives d'extraction de données. Sa puissance repose sur une combinaison d'heuristiques et de couches d'IA propriétaires permettant d'évaluer les invites et d'identifier les abus et les menaces. *« Firewall for AI protège contre le déni de service par modèle et la divulgation d'informations sensibles, grâce à des outils et des fonctionnalités disponibles pour tous les clients de l'offre WAF »*, a encore déclaré le PDG de Cloudflare. Firewall for AI exécute également une série de détections, afin d'identifier les tentatives d'injection rapide et d'autres abus, en s'assurant par exemple que le sujet reste dans les limites définies par le propriétaire du modèle.

Une IA défensive à même de détecter les comportements anormaux

Dans le cadre du programme Defensive AI, Cloudflare travaille sur des systèmes IA capables d'examiner les modèles de trafic de clients spécifiques et, à partir de là, de construire une base de référence de comportements normaux. A partir de cette base établie, toute anomalie dans les API, environnements, accès aux employés, courriels ou autres pourra être détectée. *« Defensive AI sert à comprendre comment les systèmes intelligents peuvent améliorer l'efficacité des solutions de sécurité »*, dit encore le PDG de Cloudflare. *« Cloudflare utilise l'IA pour augmenter le niveau de protection dans tous les domaines de la sécurité, que ce soit celle des applications, du courrier électronique ou de la plateforme Zero Trust de Cloudflare. Les modèles d'IA sont adaptés à une application spécifique, de sorte que la protection de l'API utilise des modèles différents de ceux du Zero Trust ou du courrier électronique »*, a-t-il encore ajouté. Quand bien même la mise en œuvre peut différer, les concepts généraux sont similaires.

Les LLM, des cibles privilégiées pour les cybercriminels

Une enquête récente a mis en lumière que seulement 25 % des dirigeants se sentiraient prêts à affronter les risques associés à l'IA. La protection des LLM représente donc un défi de taille, notamment parce qu'il est quasi impossible de restreindre les interactions des utilisateurs avec de tels systèmes dès leur conception. Ces modèles, de nature non déterministe, sont susceptibles de générer une multitude de résultats variés à partir d'un même ensemble de données. C'est pour cela que les LLM sont exposés à des risques de manipulation, de détournement et d'attaques, en faisant des cibles privilégiées pour les cybercriminels. Le système de Cloudflare propose en outre la possibilité de bloquer automatiquement ce type de menaces, sans qu'une intervention humaine soit nécessaire. Le fournisseur de cloud peut faire profiter de l'immense couverture de son réseau mondial, avec plus de 250 points de présence et un service qui peut être activé au plus près des utilisateurs finaux. Cela garantit généralement une réponse immédiate et efficace face aux attaques. De plus, Cloudflare assure une protection par défaut, sans frais supplémentaires, à tous les utilisateurs exploitant des LLM au sein de l'environnement Workers AI. Cette extension est essentielle pour réduire les risques associés à l'injection de commandes malveillantes et prévenir les fuites de données.

Détection basée sur l'apprentissage automatique

Le pare-feu WAF de Cloudflare s'appuie sur le machine learning pour bloquer automatiquement les menaces émergentes en temps réel. Il s'intègre aux autres produits de sécurité des applications de Cloudflare. Aucune formation, ni aucun service professionnel n'est nécessaire pour l'utiliser. Il se configure en seulement quelques clics. □

T.T

FONCTIONNEMENT

En plus des règles de l'OWASP, les règles gérées de Cloudflare proposent une protection rapide contre les attaques zero-day. Des ensembles de règles personnalisées permettent aux entreprises d'adapter leur pare-feu WAF afin de mettre en place des politiques totalement spécifiques à leur organisation. Le WAF s'exécute sur son réseau mondial et se place en amont des applications web afin d'être capable d'arrêter une vaste gamme d'attaques en temps réel grâce à de puissantes règles prédéfinies, de mesures de vérification des identifiants exposés, de mesures avancées de contrôle du volume des requêtes, de services d'analyse du contenu importé et de bien d'autres mesures de sécurité préétablies.

Le pare-feu WAF s'exécute sur le réseau mondial de Cloudflare en se plaçant en amont des applications web afin d'arrêter des attaques de toutes sortes en temps réel

Observabilité

Datadog déroule sa stratégie

Spécialiste de l'observabilité des données cloud, la société fondée à New York par deux Français entend capitaliser sur son cœur de métier historique, pour offrir une proposition de valeur dans d'autres domaines comme la cybersécurité, les LLM et la conformité.

Le 4 mars dernier, Datadog organisait son sommet européen annuel au Convene Sancroft, à deux pas de la cathédrale Saint-Paul, un quartier chic et touristique au cœur de Londres. C'était la troisième fois que la société posait ses quartiers dans la capitale britannique, le temps d'une journée mêlant présentations et ateliers pour les clients du spécialiste de l'observabilité des données.

L'occasion pour l'entreprise de rappeler ses fondamentaux, sur lesquels elle entend désormais construire de nouvelles solutions adaptées aux défis du moment. Lors de sa dernière conférence DASH, Datadog a en effet annoncé un panel de solutions qui sortent de son cœur de métier historique, en particulier autour des LLM, de la cybersécurité et de l'automatisation IT. Une stratégie qui demeure toutefois cohérente avec son cœur de métier initial, selon Yriex Garnier, vice-président des produits chez Datadog. *« Notre aventure a commencé autour de l'observabilité de données, et c'est autour de cette expertise fondamentale que nous étendons notre portefeuille. Car dès lors qu'on excelle dans l'observabilité, on a une visibilité sur tout le stack du système d'information. On peut donc extraire les données au service d'autres cas d'usage, de la cybersécurité au DevOps, en passant par la gestion des logs. »*

L'occasion aussi pour Datadog de donner la parole à ses clients pour illustrer la façon dont ils utilisent sa suite de produits. *« Lorsque nous lançons un produit, il arrive que nous voyions notre trafic augmenter d'un facteur vingt en quelques secondes, ce qui risque naturellement d'affecter la qualité de notre service. La solution d'observabilité de Datadog nous permet d'analyser les données issues d'événements précédents pour mieux prédire la façon dont le trafic va évoluer cette fois-ci »*, a par exemple expliqué Ana Pasparan, chargée de la gestion des plateformes cloud chez Vodafone.

Débuguer les LLM

Contexte oblige, l'entreprise a notamment profité de l'occasion pour mettre en valeur ses nouvelles solutions d'observabilité dédiées aux LLM, annoncées lors de sa conférence annuelle DASH l'an passé. L'objectif : mieux comprendre le fonctionnement des grands modèles de langage pour aider à débuguer l'applicatif, qu'il s'agisse des incohérences ou propos inappropriés dans les réponses d'un chatbot, d'hallucinations ou encore de failles de cybersécurité susceptibles d'être exploitées par des acteurs malveillants. Toute la difficulté réside dans le fonctionnement d'un LLM, qui, lorsqu'une requête lui est soumise, fait appel à différents modèles, utilise des données issues de différentes sources, fait du RAG et du fine tuning de ces données. L'observabilité vient, selon Yriex Garnier, apporter une solution à l'absence d'outils existants pour débroussailler cette complexité.

« Ce qui se passe souvent, c'est que les ingénieurs des données créent un modèle qui fonctionne très bien dans leur environnement, mais lorsque ce modèle est mis au service d'un chatbot, d'un assistant ou d'une application, ça devient beaucoup plus compliqué d'identifier d'où proviennent les bugs, parce qu'on a affaire à des boîtes noires, étant donné la complexité de la chaîne applicative impliquée. Nos outils permettent d'avoir une vue exhaustive de cette complexité pour ensuite venir débuguer les modèles. »

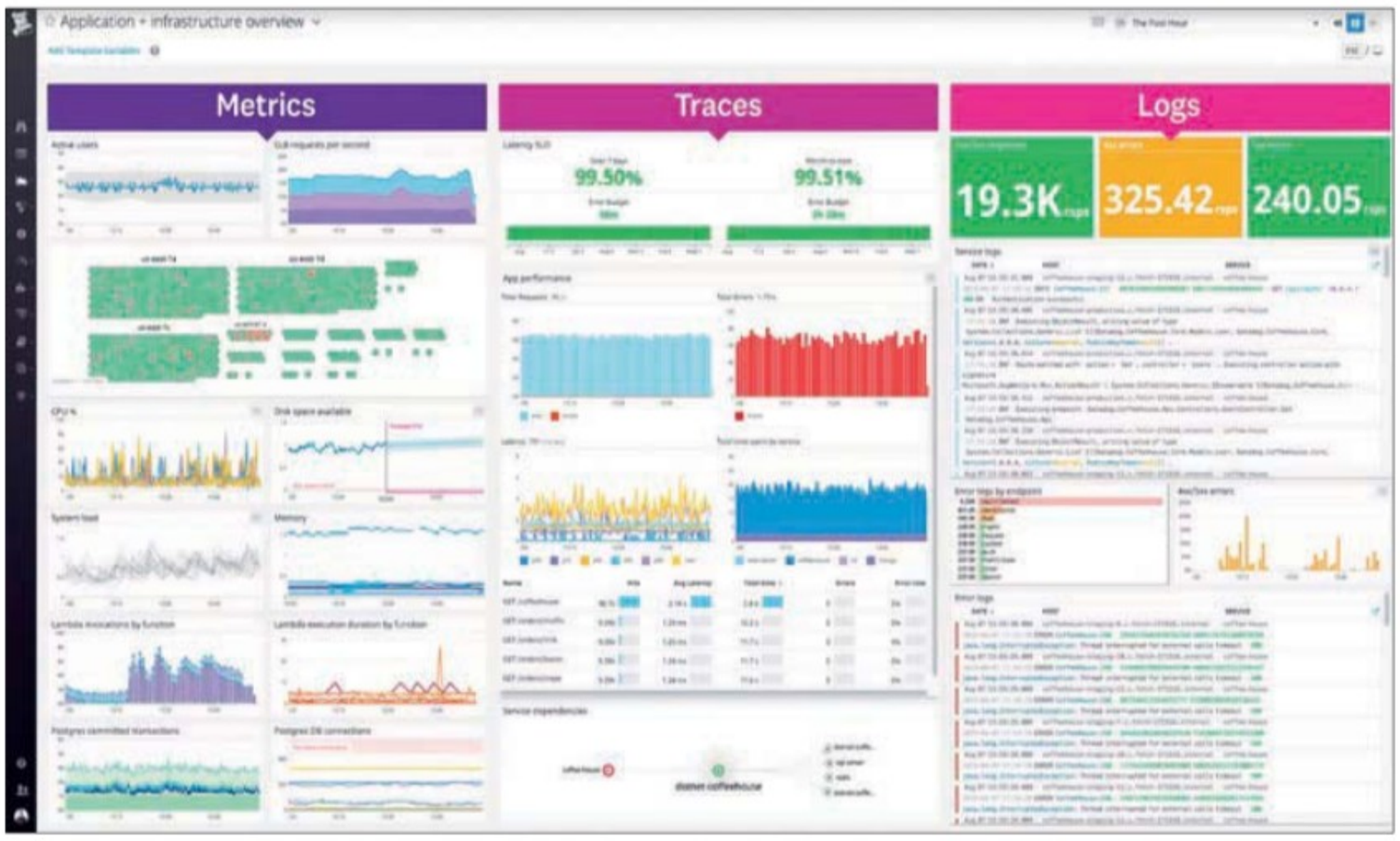
Prendre en compte le caractère évolutif des LLM

Un autre défi réside dans l'évolution des grands modèles de langage, qui tendent à se dégrader au fur et à mesure de leur utilisation. Ces modèles évolutifs doivent en effet brasser une quantité croissante d'informations à mesure

Jeremy Garcia lors de son leynote sur le Datadog Summit de Londres

qu'ils sont utilisés, ce qui finit généralement par entraîner une qualité des réponses dégradée, nécessitant de réentraîner et mettre à jour le modèle.

« Or, cette dégradation, c'est quelque chose qui est subtil et complexe à traquer : on ne peut pas se contenter d'analyser le temps de réponse et se dire que tout va bien dans la mesure où celui-ci reste stable, par exemple. Notre approche APM (Application Performance Management) nous permet à cet égard de vraiment creuser pour bien comprendre le contexte : on va, par exemple, colorer chacune des questions et réponses, des prompts, en disant "ça, c'est une question qui est spécifique à un LLM, ça, c'est une question qui est sur une application complètement différente, voilà la réponse..." On va ainsi décomposer toute la chaîne de responsabilité pour retracer le cheminement de l'information. »



Une vue de l'outil d'observabilité de Datadog.

Pour l'avenir, l'entreprise affirme plancher sur une solution permettant de sélectionner le meilleur LLM pour un cas d'usage précis, en fonction de critères de son choix, qu'il s'agisse du coût, de la performance ou encore de la qualité des réponses.

L'observabilité au service de la cybersécurité

Datadog s'efforce depuis plusieurs années d'étoffer son portefeuille de solutions autour de la cybersécurité, notamment via l'acquisition de jeunes pousses : Sqreen (sécurité des applications) en 2021, Hdiv (spécialisée dans la détection des comportements applicatifs anormaux et des vulnérabilités zero day) en 2022, ou encore Seekret (observabilité des API) la même année.

« La première étape de la sécurité, c'est de pouvoir voir ce qui se passe. Notre expertise dans l'observabilité nous permet de mettre facilement en place une surcouche d'analyse sur la visibilité des données au service de la cybersécurité », explique Max Hédouin, chef de produit chez Datadog et ancien de la jeune pousse Sqreen.

Combiner son expertise historique autour de la surveillance des environnements cloud avec l'automatisation permise par l'IA constitue la clef pour limiter le risque posé par les cyberattaques dans les environnements complexes, selon l'expert. *« Face au nombre d'alertes cyber et aux volumes de données que les analyses doivent aujourd'hui trier, une certaine fatigue peut rapidement s'installer. De premières analyses faites par des agents virtuels permettent d'enlever pas mal de contexte répétitif et d'aller directement aux attaques les plus sérieuses, qui combinent sophistication de la part des attaquants et empilement de plusieurs problèmes de sécurité pour les laisser entrer. »*

En février, les experts de Datadog ont identifié et résolu une vulnérabilité des Amazon Machine ID (AMI) susceptible de permettre aux acteurs malveillants d'exécuter du code dans des milliers de comptes AWS.

Une gestion des logs sur site pour répondre aux problématiques de souveraineté des données

Début mars, Datadog a également racheté la jeune pousse Quickwit, éditeur d'un moteur de recherche de logs open source. Ce rachat va permettre à l'entreprise d'étendre son expertise au-delà du cloud pour répondre aux besoins des clients qui doivent stocker une partie de leurs données sur site pour des questions de sécurité ou de respects des régulations en vigueur. Malgré l'essor du cloud, le stockage sur site continue en effet de faire de la résistance, et devrait continuer à s'imposer dans un monde, où les questions de souveraineté numérique et de dépendance face aux géants américains du cloud soulèvent toujours plus de questions, et où les régulations encadrant la gestion des données se multiplient.

« Nous comptons un nombre croissant de clients qui, pour X raisons, ne souhaitent pas envoyer leurs données à l'extérieur. L'idée derrière l'acquisition de Quickwit, c'est de permettre à ces clients de gérer leurs logs avec la même efficacité que celle qu'on a sur le cloud, mais dans leur propre environnement. C'est un petit peu le principe de "cloud prem" : le cloud, mais on premise (sur site) », résume Yriex Garnier. □

G.R

Stratégie

Alstom optimise ses performances cloud et HPC grâce à Oracle Cloud Infrastructure

Présent dans plus de 70 pays, Alstom a, depuis deux ans, engagé une collaboration stratégique avec Oracle pour répondre à ses enjeux d'optimisation des bases de données et de performance en calcul haute densité.

« *Nous avons vraiment suivi le chemin classique, avec une approche progressive de l'adoption du cloud chez Oracle* », explique Stéphane Detruiseux, VP Information Security et Infrastructure d'Alstom. Après une première vague d'applications migrées dans le cloud, le géant français du ferroviaire s'est rapidement confronté à une problématique de montée en charge de ses bases de données. « *Comme beaucoup, nous avions du mal à concilier montée en puissance des bases de données, maîtrise des coûts et gouvernance* », raconte le dirigeant.

Historiquement engagé avec Microsoft Azure pour ses besoins d'infrastructure, le groupe a trouvé dans Oracle Cloud Infrastructure (OCI) un complément technologique naturel, conciliant puissance, flexibilité et maîtrise de la donnée. Alstom a donc choisi de tirer parti du partenariat entre Oracle et Microsoft en s'appuyant sur les offres Exadata hébergées, afin de constituer progressivement une stack de données cloud capable de servir ses différentes régions. « *Avec Oracle, nous avons pu ajouter de la valeur de manière progressive, sans devoir revoir en profondeur nos architectures applicatives existantes* », précise Stéphane Detruiseux.

Interopérabilité multicloud et montée en puissance du HPC

Alstom a ainsi amorcé une stratégie de remise en plateforme de ses anciennes applications Oracle, migrées sur OCI tout en conservant leur logique applicative. « *C'était la solution la plus simple : nous n'avons pas eu à trop les modifier pour les adapter au cloud* », ajoute le responsable.

L'approche d'Alstom repose également sur l'interopérabilité renforcée avec les autres hyperscalers. Grâce aux accords intercloud noués par Oracle, l'industriel a pu déployer des stacks OCI chez plusieurs fournisseurs de cloud pour répondre aux spécificités de ses marchés régionaux. « *Nous avons commencé à intégrer des cas d'usage en IA générative sur certains projets, avec toutefois des difficultés à ajuster les capacités au bon rapport qualité/prix. Dans ce domaine, OCI s'est révélé le plus adapté et le plus conforme à nos attentes* », souligne-t-il.

Côté ingénierie, OCI est également utilisé pour accélérer les simulations HPC (High Performance Computing). « *Nos ingénieurs conçoivent des trains soumis à des contraintes très complexes à modéliser. Certains scénarios nécessitaient jusqu'à plusieurs mois de calcul. Avec OCI, nous sommes passés à quelques semaines, voire quelques jours, selon les cas. Cela nous permet aussi d'explorer des scénarios que nous écartions auparavant, faute de puissance disponible* », explique Stéphane Detruiseux.

Cyril Grira, VP Cloud & AI Oracle France et Stéphane Detruiseux, CISO & CTO Alstom lors de l'Oracle CloudWorld Tour de Paris

Une gestion localisée et conforme des données

Dans un contexte où la souveraineté numérique devient un enjeu stratégique, la localisation des données est également au cœur des préoccupations d'Alstom et de ses clients. « *Nous avons eu un cas très concret en Arabie saoudite, où un client exigeait que les données soient hébergées localement. Grâce à OCI, qui dispose d'infrastructures sur place, nous avons pu déployer les bases de données dans la région concernée, tout en maintenant le reste de la stack dans d'autres environnements cloud* », détaille-t-il.

Cette architecture multicloud maîtrisée est opérée par BMC, partenaire d'Alstom, qui pilote l'ensemble des environnements techniques. « *Ce modèle nous apporte la flexibilité dont nous avons besoin, tout en respectant les exigences de nos clients en matière de conformité, de localisation et de performance* », conclut Stéphane Detruiseux. □

O.B

Cybersécurité

Une construction permanente

Pharaonique, le projet ITER sur la fusion atomique reposera à terme sur 200 réseaux industriels. Sécuriser l'IT comme l'OT de ce SI en travaux repose sur un mix d'outils classiques et d'approches originales.

Organisation internationale et projet recherche œuvrant sur la fusion nucléaire, initié en 1985, et lancé à la suite d'un accord signé en 2006 entre six grands pays (États-Unis, Chine, Russie, Inde, Japon, Corée) et l'Europe, ITER ambitionne de reproduire et maîtriser l'énergie, à l'instar de celle alimentant le soleil. Un objectif qu'une partie des scientifiques spécialisés jugent illusoire mais annoncé par d'autres et bien sûr, par les promoteurs d'ITER, à l'horizon 2040. Concrètement, les promoteurs du projet espèrent ouvrir la voie à des prototypes de réacteur destinés à produire de l'électricité à l'échelle industrielle. Le projet est financé à hauteur d'environ 50 % par l'Europe et d'un peu moins de 10 % pour les autres membres. Le plus gros de ces contributions n'est pas financier, mais prend la forme d'équipements industriels et scientifiques hors normes. Les pays se partagent tous les résultats expérimentaux et les données scientifiques.

Au cœur de la Provence, ITER ambitionne de dompter la fusion nucléaire, l'énergie des étoiles.

En attendant la mise en route, la construction du site et des différents systèmes industriels nécessaires se poursuit, et implique parallèlement la construction d'un système d'information et sa sécurisation. Les spécificités du projet se concrétisent par autant de particularités dans le SI. Romain Bourgue, RSSI, illustre : *« À terme, le site comptera autour de 200 réseaux industriels, provenant de différents pays, pour opérer le réacteur. Avant de les intégrer, nous analysons chacun d'entre eux d'abord sur le plan de la sécurité OT. »* Les réseaux et équipements intégrés remontent près d'un million de « process variable » (pression, température…). *« Nous vérifions l'interface de ces réseaux avec le système de contrôle centrale »*, ajoute Romain Bourgue.

Les volumes de données scientifiques générées sont également hors norme. À terme, ce sont plusieurs pétaoctets (1015 octets) par jour qui seront générés par ITER. Ces données doivent être archivées et dupliquées à des fins de sécurité, mais aussi rester disponibles pour les pays partenaires. *« Le Japon possède d'ores et déjà une salle et une connexion lui donnant un accès et une visualisation en temps quasi réel à ces données »*, détaille Romain Bourgue.

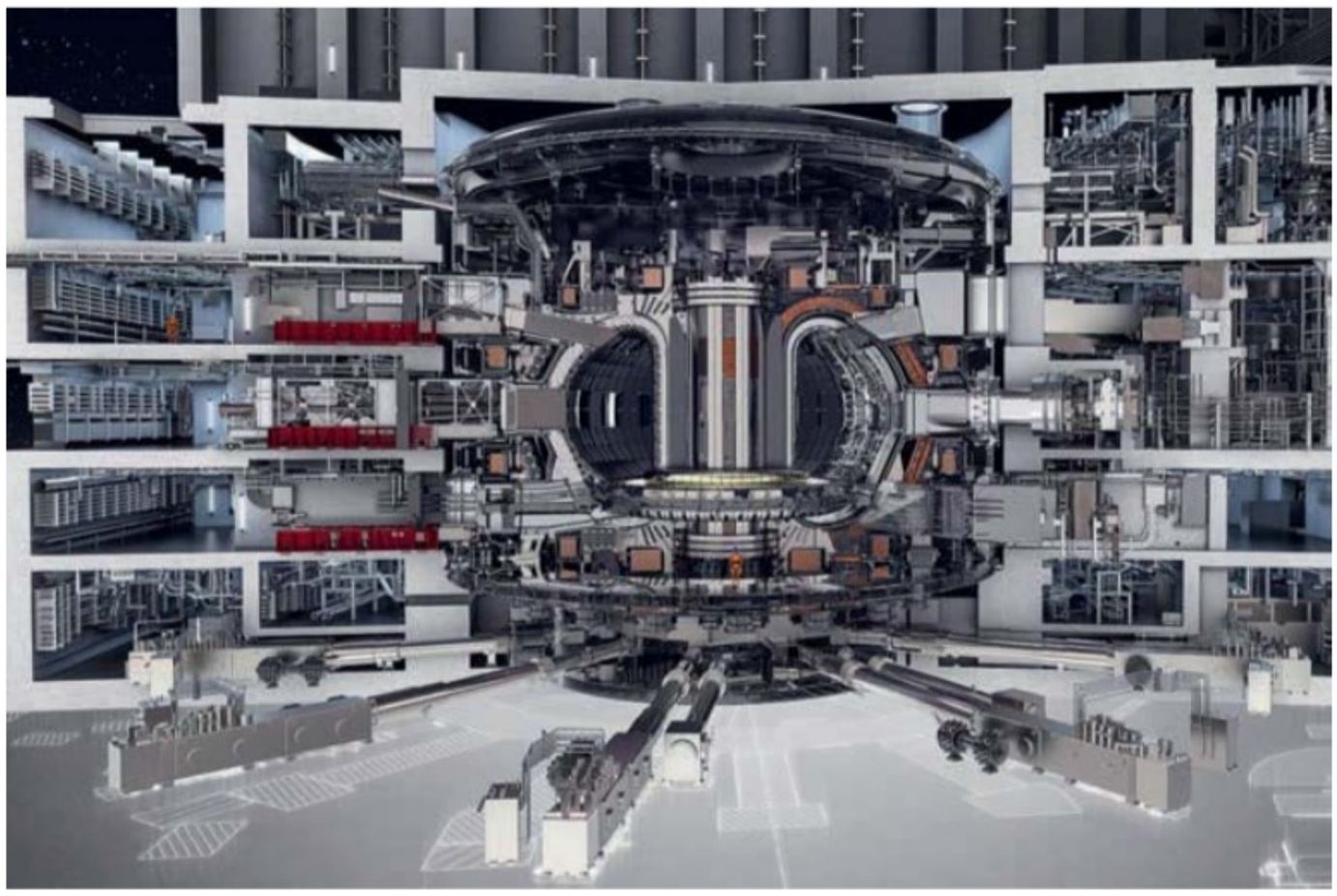
Côté infrastructure, le SI est relié à une salle blanche de Digital Realty (ex-Interxion) via une fibre de 50 km. Ce datacenter tiers supportera à terme un PRA pour le SI. *« Nous profitons également chez ce fournisseur d'une interconnexion direct avec nos partenaires pour faciliter l'accès à nos données »*, ajoute notre interlocuteur. Le site d'ITER compte également depuis peu un troisième datacenter local dédié au calcul scientifique et doté à terme de 17 000 cœurs. *« L'idée est d'offrir aux partenaires des possibilités de calcul avancé pour travailler directement les données »*, explique Romain Bourgue.

LE MAINTIEN DU PLASMA DANS LE TEMPS

Le 17 février dernier, le CEA, qui opére le réacteur Tokamahk de Cadarache, a réussi à maintenir un plasma pendant 22 minutes, un record dépassant le temps annoncé par la Chine en janvier. Ce plasma est constitué d'atomes auquels les électrons ont été enlevés, ce qui passe par une montée en température de l'ordre de 100 millions de degrés Celsius. L'institution a indiqué que ce record indique une maitrise à la fois dans la production, mais aussi dans le maintien du plasma. Une étape cruciale sur le chemin de la fusion semble être franchie. Des résultats qui vont également impliquer un niveau élevé de cybersécurité.

Une sécurisation complexe

Pour assurer la cybersécurité de l'IT comme de l'OT, Romain Bourgue et son équipe de quatre personnes, ont mis en place des moyens classiques. *« Nous faisons face aux mêmes menaces, et nous utilisons les mêmes outils et les mêmes méthodes que les autres organisations »*, résume-t-il. Côté outils : un SIEM traite 7000 évènements par seconde, un SOC interne et des outils de sécurisation classiques, reverse proxy... *« Après avoir fait le choix de solutions internes customisées, nous sommes en train d'évaluer la migration sur le cloud (Sentinel) d'une partie du SIEM. MS Defender est également en production pour compléter la protection du SI. Globalement, ce qui a du sens pour nous demeurera en interne, le reste sera externalisé »*, détaille le responsable. L'équipe cyber utilise également des sondes chargées de monitorer l'opérationnel et de *« détecter la présence d'un Windows 7 pas patché depuis 15 ans sur un équipement industriel »*, détaille Romain Bourgue. L'OT demeure difficile à sécuriser, entre autres parce qu'il faut tenir compte de l'obsolescence. Autre facteur sensible, les réseaux hétérogènes, par nature, nécessitent une architecture de sécurité adaptée. Le défi est donc de trouver l'équilibre entre ouverture et cloisonnement. *« La sécurité du SI s'appuie sur un principe de défense en profondeur, avec des protections adaptées à la criticité de chaque réseau. Les flux d'information sont strictement contrôlés : ils peuvent descendre vers des niveaux moins sensibles, mais jamais remonter vers des zones plus critiques »*, décrit Romain Bourgue. En cas de problème, il reste la possibilité de couper tout ou partie des liens avec internet. Autre difficulté, la gestion des identités est complexe. 8 000 contributeurs issus de 33 pays différents participent au projet et doivent s'authentifier sur le SI. *« L'arbitrage entre sécurité et utilisabilité est souvent délicat »*, souligne notre interlocuteur.



Pièce maîtresse d'ITER, le plus grand tokamak au monde : un réseau de bobines supraconductrices génère un champ magnétique confinant un plasma à 150 millions de degrés. Objectif : maîtriser la fusion nucléaire pour produire 500 MW d'énergie propre.

Romain Bourgue, RSSI.

Aujourd'hui, sécuriser tous les composants de ce projet titanesque constitue un défi de taille que relève Romain Bourgue et son équipe dans une démarche presque "expérimentale". La nature internationale du projet ITER reste néanmoins une force : *« Nous travaillons tous les jours avec nos collègues de 33 pays du monde. Cette collaboration dépasse les frontières et les problématiques étatiques. »* Avant d'ajouter, *« pour une organisation internationale comme ITER, la souveraineté n'est pas un sujet »*.

Il s'agira de parer à de nouvelles menaces comme de découvrir et sécuriser de nouveaux équipements. Point de satisfaction, l'usine de cryogénie, qui assurera le refroidissement des bobines du réacteur à -269 °C, a validé ses premiers tests. *« Elle est unique par sa puissance et intéresse beaucoup de monde, en particulier, pour les applications quantiques »*, se réjouit Romain Bourgue. En définitive, la cybersécurité d'ITER rappelle que protéger infrastructures critiques d'un projet international exige une adaptation permanente, où l'anticipation des risques numériques et la coordination entre des acteurs très divers restent un équilibre fragile à préserver. □

P. Br

Industrie Audi virtualise sa production

Le constructeur automobile rend opérationnel son projet d'automate programmable virtualisé en s'appuyant sur le cloud de Broadcom.

Edge Cloud 4 Production (EC4P) d'Audi, reposant sur le logiciel VMware Cloud, est désormais opérationnelle. Le premier automate programmable virtuel (vPLC) a été déployé dans l'usine de Boellinger Hoefe, en Allemagne, où Audi produit son modèle électrique e-tron GT. Avec EC4P, Audi centralise la gestion et la maintenance de ses équipements informatiques industriels, simplifie l'application des correctifs de sécurité, et réduit son impact environnemental en limitant l'usage de matériel physique et les interventions manuelles.

Un nouveau mode d'automatisation

Audi transforme ses usines en sites de production intelligents, en intégrant une automatisation pilotée par logiciel directement sur la ligne de production et en rapprochant les mondes de l'IT et de l'OT (Operational Technology). Cette initiative repose sur une collaboration étroite avec plusieurs partenaires technologiques clés, dont Broadcom, Cisco et Siemens. Ce projet est un élément du programme 360factory, qui vise à rendre la production plus efficace et conduite par les données. A terme, ce cloud privé sera étendu à l'ensemble des sites de production, afin de tirer parti des avancées du contrôle numérique dans les processus industriels du constructeur.

Une consolidation matérielle d'ampleur

Avec EC4P, le constructeur réduit son empreinte matérielle en remplaçant des milliers de PC industriels décentralisés par une architecture plus efficace, évolutive et flexible, reposant sur des serveurs en périphérie locaux qui unifient le cloud et la production. Ainsi, VMware Cloud Foundation a été déployé pour créer un environnement de cloud privé à l'extérieur de l'usine de Boellinger Hoefe, où les charges de travail critiques de production sont hébergées et gérées de manière centralisée. Au lieu de maintenir des PC industriels physiques pour des milliers de postes de travail sur la ligne de production, ces derniers fonctionnent désormais sous forme de machines virtuelles (VM) exécutées sur VMware Cloud Foundation, en dehors de l'usine. Les mises à jour des logiciels et des systèmes d'exploitation peuvent être effectuées simultanément, sans perturber les changements de poste. En cas de dysfonctionnement, un poste de travail virtuel peut être remplacé instantanément à distance.

Des automates virtuels

Les automates pilotent les robots qui assemblent différentes parties des véhicules. Hébergés sous forme de machines virtuelles ou de conteneurs, ils sont gérés comme n'importe quelle infrastructure cloud IT. Les mises à jour de configuration, les correctifs de sécurité et les évolutions fonctionnelles peuvent être déployés directement depuis le cloud privé d'Audi.

Audi automatise virtuellement ces sites de production avec le projet EC4P.

L'usine du futur

Audi explore également de futurs cas d'usage, tels que l'intelligence artificielle appliquée à la production, l'analyse de données et la vision par ordinateur. Grâce à VMware Cloud Foundation, Audi vise à atteindre plusieurs objectifs majeurs à l'usine de Boellinger Hoefe, comme la mise en place d'une plateforme de cloud privé unique pour l'ensemble des applications industrielles assez flexible, permettant d'adapter rapidement les lignes de production aux évolutions du marché. Par ailleurs, cette évolution apporte de nombreux gains en termes de réduction du matériel utilisé et sur la sécurité et l'empreinte carbone. Les coûts sont diminués par la réduction du nombre de matériels, avec une maintenance simplifiée et une centralisation des mises à jour logicielles. L'empreinte matérielle plus faible a pour conséquence une consommation d'énergie optimisée et moins de déchets électroniques. La correction des vulnérabilités à grande échelle et la restauration rapide en cas d'attaque garantissent la continuité de production, ainsi qu'une surveillance proactive des charges de travail et une maintenance automatisée pour minimiser les temps d'arrêt. □

B.G

L'intelligence artificielle en 30 questions

La Documentation française fournit depuis des lustres des contenus de référence dans de nombreux domaines. Récemment elle a sorti plusieurs ouvrages sur l'intelligence artificielle dont celui-ci. Le livre répond aux principales questions que peuvent se poser les gens sur l'intelligence artificielle. L'ouvrage livre des réponses claires et définit clairement les différents concepts ou technologie qui se rapportent à l'intelligence artificielle. Au moment où tout est dit sur cette technologie, il nous a semblé bon de revoir ces notions de base. Réalisé pour le grand public, cet ouvrage donne toutes les clés d'analyse pour comprendre ce qu'est l'intelligence artificielle, son fonctionnement et ses applications dans la vie quotidienne, ainsi que les limites ou les enrichissements qui doivent lui être apportés. L'objectif du livre est de permettre à un public non spécialiste de comprendre les enjeux entourant l'intelligence artificielle en fournissant des outils pédagogiques de compréhension.

Le point sur

Entre le moment où nous écrivons ce livre et celui où vous le lirez, les géants de la « Tech » auront certainement présenté de nouveaux systèmes d'IA qui ridiculiseront le GPT 4o, révélé par OpenAI en mai 2024. Pourtant, cette IA, capable de discuter avec une voix très expressive de tous les sujets et d'interpréter des vidéos relevait, il y a peu, de la science-fiction. Si les techniques évoluent à grande vitesse, les principes que nous présentons ici resteront des repères utiles pour mieux comprendre cette formidable technologie.

Une technologie de pointe à la portée de tous

Ces dernières années, l'intelligence artificielle était dans tous les esprits, en grande partie à cause des fantasmes qu'entretiennent des œuvres de fiction autour de la prise du pouvoir par les machines mais aussi parce que, dans l'industrie et la recherche, les progrès du machine learning, des « réseaux de neurones » et du deep learning, donnaient accès à des résultats jusque-là inespérés dans divers domaines : l'exploration spatiale, l'interprétation des images médicales, le déchiffrement de manuscrits anciens, l'amélioration de la précision des prévisions météorologiques. Toutes ces applications semblaient concerner des systèmes complexes pour lesquels l'informatique permettait au cerveau humain de résoudre des problèmes qui dépassaient ses capacités. L'IA était un outil sophistiqué, manipulé par des spécialistes pour traiter des problèmes très pointus.

La révélation de ChatGPT, en novembre 2022, a changé la nature de l'engouement pour cette technologie. En réalisant que l'IA « parlait » comme nous, comprenait (quasiment) comme nous, le grand public a pris conscience des capacités de l'IA. Curieusement, c'est quand l'IA s'est mise à faire ce que chacun sait faire (comprendre des requêtes exprimées en langage naturel et les exécuter) que le plus grand nombre a évalué le potentiel de cette technologie pourtant quasiment octogénaire. Du fait de sa grande facilité d'utilisation et de la capacité de chacun à juger de la pertinence de ses réponses, ChatGPT a permis à l'IA d'acquérir la notoriété des autres grandes innovations technologiques de notre histoire : le feu, la roue, l'électricité, Internet.

Pourtant, avant même l'arrivée de ChatGPT, l'intelligence artificielle était déjà omniprésente dans notre vie mais, à l'instar de l'électricité, nous ne la manipulions pas en tant que telle mais accédions à des services qui eux-mêmes l'utilisaient. Les moteurs de recherche sur Internet, les filtres antispams de nos boîtes courriel, les systèmes de recommandation de produits sur les sites de vente en ligne, les logiciels de navigation dans nos véhicules, la reconnaissance vocale qui permet de dicter des SMS à haute voix, la reconnaissance de visage qui déverrouille nos smartphones… tous ces services du quotidien, avec lesquels nous vivons depuis des années, reposent sur des algorithmes d'IA. ChatGPT n'est donc qu'une nouvelle expression de l'IA, mais elle est mise à notre disposition sans véritable usage. A priori, elle ne fait qu'écrire des textes en réponse à des demandes. Chacun peut utiliser cette capacité brute à en faire ce qu'il veut : écrire du code informatique, résumer des articles, composer des chansons, comparer des textes… C'est un peu la même différence qu'entre donner un poisson à quelqu'un ou lui enseigner la pêche.

Un des premiers événements qui a fait réaliser au grand public l'avènement de l'intelligence artificielle est la défaite de Garry Kasparov, champion du monde d'échecs, lors de son match contre l'ordinateur Deep Blue d'IBM, en 1997. Le jeu d'échecs est communément reconnu comme requérant une grande intelligence. Le fait qu'un ordinateur puisse battre un champion d'échecs était le signal que, au moins pour cet aspect de l'intelligence, la machine était plus performante que l'homme. Sentant bien que cette conclusion avait quelque chose de mortifère pour l'avenir de l'humanité, une partie des personnes bien informées affirmèrent que le jeu d'échecs était finalement assez simple et que la machine n'avait pas eu beaucoup de mérite à gagner. Ils ajoutaient qu'avec un jeu vraiment compliqué et stratégique, comme le go, cela n'arriverait jamais. En 2017, le programme AlphaGo de DeepMind devient champion du monde de go, battant les meilleurs joueurs du monde, en jouant des coups improbables et même déconseillés dans les écoles de go. La conclusion de cette escapade de l'IA dans le monde des jeux de société est qu'il vaut mieux éviter de dire que quelque chose est impossible avec l'IA.

De l'IA spécialisée à l'IA multimodale et polyvalente, voire générale ?

Une limitation rassurante souvent avancée est celle de la spécialisation de l'IA. Une IA entraînée à reconnaître des images de chatons ne saurait reconnaître des papillons. Une IA spécialisée dans la vision ne saurait être compétente dans le traitement du langage. Cette présomption fut longtemps exacte. Elle l'est moins aujourd'hui. L'apparition de très larges modèles d'IA utilisant des centaines de milliards de paramètres, entraînés de nombreuses fois sur des jeux de données colossaux, rend l'IA plus robuste et polyvalente. L'IA est aussi devenue multimodale (combinant texte, image, audio) et interopérable, ce qui décloisonne ses applications et ses possibles.

LES SAISONS DE L'INTELLIGENCE ARTIFICIELLE

L'IA est née dans les années 1950 et les principes qu'elle se proposait de développer étaient prometteurs. Les militaires attendaient notamment des applications essentielles comme la traduction automatique. Mais les techniques de l'époque ne permettaient pas d'atteindre les résultats espérés. Les financements se sont taris et, entre 1974 et 1980, les développements dans le domaine de l'IA se sont quasiment arrêtés. Ce fut le premier hiver de l'IA. Au début des années 1980, l'apparition des systèmes experts relance l'intérêt pour l'intelligence artificielle, notamment dans les entreprises. Si cette technique est toujours utilisée, elle ne se cantonne qu'à résoudre des problèmes relativement simples et ses limites inhérentes vont provoquer une nouvelle pause dans son essor. C'est son second hiver, qui durera de 1987 à 1993. Avec l'arrivée d'Internet, la multiplication des données disponibles et l'augmentation des puissances de calcul des ordinateurs, des techniques d'IA basées sur les données, qui avaient été imaginées dans les années 1950 mais n'avaient pas pu être mises en œuvre, vont pouvoir se déployer. L'ère du machine learning a commencé et l'IA connaît un été ininterrompu depuis.

Qu'est-ce que l'intelligence artificielle ?

Un concept valise

Le terme d'« intelligence artificielle » a été proposé en 1956 par John McCarthy du MIT. Son collègue Marvin Minsky expliquait que c'était *« la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique »*. L'intelligence artificielle est, depuis, devenue un concept valise qui recouvre énormément de choses. Certaines définitions très englobantes couvrent quasiment tous les logiciels alors que d'autres sont plus pointues.

Une modélisation comme aide à la décision

L'objectif est que les programmes informatiques fassent autre chose que des opérations sur des nombres (multiplications, additions, soustractions…) et qu'ils traitent et modélisent également des concepts comme le fait le cerveau humain. Dans la mémoire d'un ordinateur, ces manipulations permettent d'évaluer une situation, de prendre des décisions, et de proposer un plan d'action afin de résoudre un problème.

Des techniques pour interpréter, comprendre, traiter

En termes d'usage, on peut retenir que les techniques d'IA sont aujourd'hui les seules à pouvoir interpréter le contenu d'une image, comprendre un texte et traiter les explosions combinatoires. Même si elles peuvent faire bien autre chose, ce sont sur ces trois sujets qu'elles règnent désormais sans partage.

Questions-réponses

L'explosion combinatoire : l'exemple du jeu d'échecs

Lorsque le nombre de cas possibles à étudier devient difficile à dénombrer, on parle d'explosion combinatoire. Par exemple, au jeu d'échecs, le premier joueur a 20 coups possibles, le second joueur en a 20 également. Après le premier tour, il existe ainsi 400 configurations d'échiquier possibles. À la fin du deuxième tour, il y en aura environ 400 x 400, soit 160000, et ainsi de suite. Un programme d'échecs va donc gérer cette multitude de cas possibles avant de prendre une décision.

Régler des problèmes complexes : l'exemple de la cuisine

La différence entre un logiciel « conventionnel » et un logiciel d'IA est liée à la nature des problèmes gérés. Si l'on compare un programme classique à une recette de cuisine, le logiciel donne les instructions pour passer des données d'entrée (les ingrédients) à une donnée de sortie (le plat prêt à être mangé). C'est utile mais le plus compliqué est de répondre à la question *« qu'est-ce que je vais faire à manger ce soir ? »* C'est pour ce type de problème qu'il est possible de se faire aider par une intelligence artificielle : on peut lui demander de comparer tous les ingrédients disponibles à toutes les recettes connues dans les livres de cuisine pour en extraire celles qui sont faisables avec ce que l'on a. Ce qui n'est pas loin de relever de l'explosion combinatoire.

Une définition officielle

Le Journal officiel définit l'IA comme le *« champ interdisciplinaire théorique et pratique qui a pour objet la compréhension de mécanismes de la cognition et de la réflexion, et leur imitation par un dispositif matériel et logiciel, à des fins d'assistance ou de substitution à des activités humaines »*.

Quelles sont les différentes techniques d'apprentissage ?

L'apprentissage supervisé

L'apprentissage est dit supervisé car la réponse attendue pour chaque question posée est connue. Il y a donc un moyen de s'assurer que le réseau donne ou non la réponse correcte. L'opération qui consiste à associer la bonne réponse à la question posée est appelée l'annotation. C'est une partie très lourde de ce type d'apprentissage car il faut un grand nombre de données annotées, généralement par des humains.

L'apprentissage non supervisé

Les données annotées sont parfois difficiles à collecter, notamment quand il existe peu d'exemples de ce qu'on cherche ou qu'on ne sait pas ce que l'on cherche. L'application la plus répandue en est la constitution de groupes d'éléments similaires dans une population. L'apprentissage non supervisé consiste à identifier les caractéristiques qui permettent de constituer des groupes homogènes. À l'issue de l'apprentissage, les critères pour être associé à un groupe ou à un autre sont établis.

L'apprentissage semi-supervisé

Ce type d'apprentissage utilise des données pour lesquelles les données annotées sont très simples à construire. En effet, au lieu de chercher la réponse à une question donnée, on va ici construire la question à partir d'une réponse connue. L'apprentissage sera supervisé, puisque la question et la réponse sont connues, mais la phase d'annotation des données d'entrée, indispensable à l'apprentissage supervisé, sera économisée.

Éditeur : La Documentation française
Format physique : EAN : 9782111579231, 102 pages
Format PDF : EAN : 9782111579248, 96 pages
Format ePub : EAN : 9782111579255

Lire la suite sur www.laviepublique.fr

Questions-réponses

Reconnaître des animaux

Pour apprendre à une IA à identifier les animaux représentés sur une photo, il faut lui donner des milliers de photos d'animaux annotées. Pour la phase d'apprentissage, si l'entrée du système est une photo du cheval, la sortie est le mot « cheval ». Les données utilisées pour l'apprentissage sont des milliers de photos de chevaux associées au mot « cheval », des milliers de photos de chiens associées au mot « chien » et ainsi de suite. Après cet apprentissage, l'IA sera capable de reconnaître tout animal vu pendant l'apprentissage.

La catégorisation des clients

Chaque consommateur fréquentant les boutiques en ligne a pu expérimenter l'efficacité de l'apprentissage non supervisé. Lorsque, à l'issue d'un achat, des recommandations pour d'autres produits nous sont faites, c'est parce que nous avons été identifiés comme faisant partie d'un groupe de clients ayant les mêmes appétences pour certains articles. L'IA identifie, sans supervision d'un expert du marketing, un groupe de clients qui se ressemblent parce qu'ils ont acheté le dernier Astérix, un disque de Stromae et un grille-pain.

Boucher les trous dans les images

Sur les téléphones portables, une application permet de retirer un baigneur inconnu d'une belle photo de coucher de soleil sur une plage. L'IA remplit le vide avec le bout de plage, de mer et de ciel rougi. L'apprentissage d'un tel système est fait de façon semi-supervisée. On retire une partie d'une image de paysage. L'image « percée » est utilisée comme entrée et l'image intacte est utilisée comme vérité terrain. Entraînée avec des millions d'exemples comme celui-ci, l'IA sera capable de boucher les trous dans toutes les images.

Quels sont les liens entre IA et robotique ?

Le robot pour compenser les faiblesses de l'IA

L'une des plus grandes déficiences de l'IA est son rapport au monde. L'utilité et la valeur ajoutée de l'IA ne se font ressentir que par ses interactions avec son environnement. Dans le même temps, elle est dépourvue d'un corps qui lui faciliterait ces interactions. Le robot est donc le vecteur idéal par lequel elle peut agir et évoluer dans le réel. La mobilité, la capacité de détecter un obstacle, à prendre ou à déplacer des objets sont des expériences lui permettant une meilleure compréhension de l'univers qui l'entoure, un apprentissage que seul un robot peut lui offrir. Cette confrontation tangible lui permet ainsi de collecter ce dont elle a le plus besoin pour évoluer et s'améliorer, la donnée.

L'IA pour compenser les faiblesses du robot

Il est toujours possible d'embarquer dans un robot de « simples » logiciels. Le robot ainsi programmé pourrait, sans IA, se mouvoir d'un point A à un point B, monter un escalier ou bien prendre une vis et la déposer dans une boîte. Les choses se compliquent si un obstacle survient durant l'action. Le programmeur peut toujours tenter d'anticiper la survenance de cet obstacle mais jusqu'à quel point ? Le salut face à une telle complexité ? L'intelligence artificielle. Sa capacité en temps réel à traiter un nombre incommensurable de données et à réagir en conséquence en fait l'outil parfait. Connectée à l'ensemble des capteurs du robot (température, mouvement…), l'IA est capable d'extraire les données d'une situation particulière, les traiter et apporter une compréhension fine et efficace au robot.

Questions-réponses

Exemples de robots dotés d'IA

- **Nao et Pepper** (conçus par la société française Aldebaran) sont des robots humanoïdes, respectivement de 58 et 120 cm de haut, capables de détecter les visages et de converser avec les hommes.
- **L'aspirateur intelligent Roomba** (société iRobot) navigue et cartographie son environnement.
- **Bras robotisé :** afin d'aider les robots à détecter des objets, apprendre puis répéter les mouvements humains.
- **Robots de livraison ou de logistique :** pour la navigation en extérieur (robots de Starship Technologies) ou en intérieur dans des commerces (Plato de la société Aldebaran) ou dans des entrepôts (robots d'Amazon Robotics).

Exemple de capteurs utilisés en robotique et en IA

L'IA peut identifier avec une bonne efficacité les objets mais aussi les hommes. À la détection d'un visage humain, le robot est capable d'engager une ou plusieurs actions. Cela peut être un simple mouvement de la main pour le saluer ou l'inviter à se présenter. Dans ce dernier cas, une IA conversationnelle prend le relais. Celle-ci recourra aux microphones (capteurs) du robot pour interpréter ce qui lui est dit. La liste des cas d'usage où robot et IA agissent de concert est infinie : de l'industrie où des robots dotés d'IA sont utilisés dans le contrôle qualité de pièces détachées, à la logistique où des robots naviguent dans d'immenses entrepôts en passant par la médecine avec des robots chirurgiens. La combinaison de ces champions technologiques devrait permettre de nombreuses évolutions en matière de véhicules autonomes, de drones, de jouets intelligents, ou de robots d'exploration spatiale…

Bioéconomie

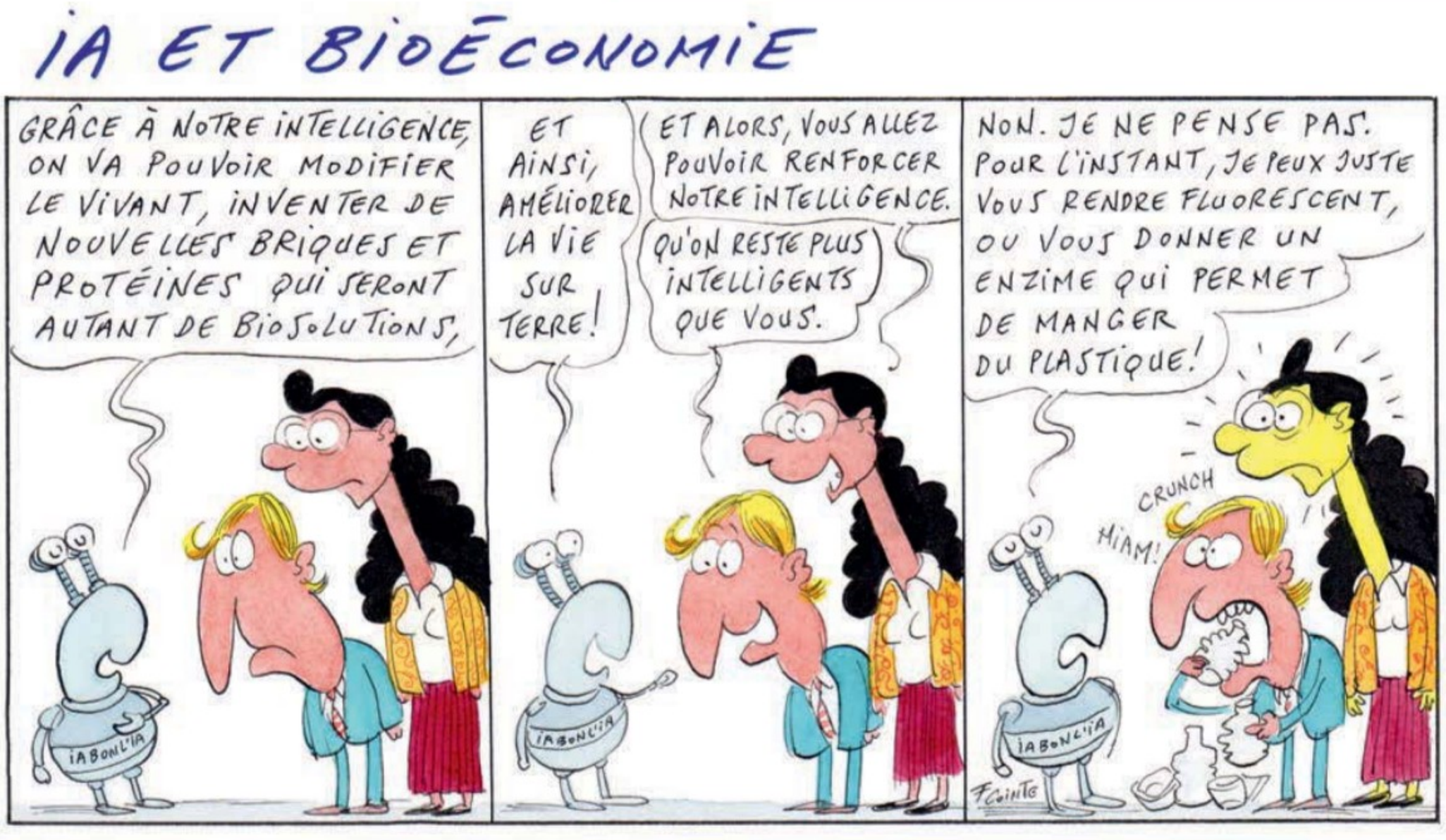
Capgemini propose une nouvelle méthodologie pour l'ingénierie des protéines

Cette méthodologie utilise un grand modèle de langage spécialisé dans les protéines (pLLM) pour identifier les variantes de protéines les plus efficaces.

Cette nouvelle approche, en passe d'être brevetée, contribuera à accélérer les progrès de la bioéconomie mondiale2 et à réaliser des avancées scientifiques clés dans des secteurs tels que la santé, l'agriculture et les sciences de l'environnement. Créée par Cambridge Consultants, elle a été appliquée à plusieurs cas d'usage clés, afin de démontrer comment elle pouvait radicalement accélérer l'innovation, et qui peuvent être facilement transposés à d'autres applications. Ainsi, l'approche de Capgemini s'appuyant sur l'IA générative a amélioré l'enzyme cutinase, augmentant de 60 % sa capacité à dégrader le plastique PET. Cette avancée illustre la manière dont l'ingénierie des protéines peut créer de nouvelles solutions très efficaces et peu onéreuses pour lutter contre les déchets plastiques dans le monde. En facilitant la dégradation du plastique, cette avancée peut soutenir les objectifs de durabilité et contribuer à réduire les coûts opérationnels liés à la gestion des déchets. De plus, en utilisant les prédictions de l'IA générative, Capgemini a réduit le nombre d'expériences nécessaires pour identifier une variante améliorée de la protéine fluorescente verte, couramment citée comme référence, de plusieurs milliers à seulement 43 points de données, atteignant un niveau de luminosité sept fois supérieur à celui de la protéine naturelle des méduses. Cela réduit considérablement le temps et les ressources généralement nécessaires aux tests expérimentaux, permettant un déploiement plus rapide dans un large éventail de domaines, qu'il s'agisse d'accélérer la découverte de médicaments ou d'améliorer les outils de diagnostic à l'avancement des applications de bio-ingénierie. □ **B.G**

UN NOUVEAU LAB IA

Avec le soutien de Capgemini, l'Institut IA et Société, l'École normale supérieure (ENS-PSL) et la Fondation de l'ENS lancent un Observatoire dédié à l'analyse des impacts environnementaux de l'intelligence artificielle (IA) à toutes les étapes de son cycle de vie (entraînement, ajustement, inférence et fin de vie) et à leur mitigation. Cet observatoire a pour ambition d'établir une méthodologie solide et partagée pour promouvoir des pratiques durables d'utilisation de l'IA.



Jumeau numérique

Au service du Pavillon France de L'Exposition universelle d'Osaka

Après la restauration de Notre-Dame, Dassault Systèmes met sa plateforme 3D à disposition du Pavillon France de l'Exposition Universelle d'Osaka. Les équipes de Dasty ont mis à la disposition du Cofrex (Comité français pour l'Exposition universelle), un téléport avec un jumeau numérique du Pavillon français pour finaliser son aménagement et proposer une expérience immersive unique à ses futurs visiteurs virtuels.

Dassault Systèmes a joué un rôle central dans la conception de l'aménagement et de la scénographie du futur Pavillon France de l'Exposition universelle d'Osaka (cf. encadré), en s'appuyant sur l'expérience acquise lors de fouilles en Égypte, pour la conservation de la Grotte de Lascaux ou la restauration de Notre-Dame de Paris. Pour cela, l'entreprise de Vélizy a rassemblé toutes les fonctionnalités de sa plateforme 3DEXPERIENCE dans un téléport 1/1, alimenté par la suite logiciels prototype Teleporteam, lequel figurait une maquette numérique à l'échelle 1 de cette future construction éphémère. *« Le téléport n'a pas vocation à se substituer à nos solutions. C'est un révélateur de la puissance de la 3D, c'est un visualisateur de projet, des capacités de notre plateforme. Il permet de résoudre le paradoxe de la 3D : de plus en plus de données 3D sont produites mais on continue de les visualiser en 2D. Le fait d'être immergé à plusieurs dans le même espace 3D change le rapport de perception et de décision »*, précise Medhi Tayoubi, vice-président Stratégie & Innovation de Dassault Systèmes, passionné d'archéologie et de patrimoine, à l'origine des collaborations transdisciplinaires de Dasty avec la mission ScanPyramid en Égypte, dont il est co-directeur, ou La Cité de L'Architecture et du Patrimoine.

Pour Jacques Maire, le commissaire général du Cofrex (Comité français pour l'Exposition universelle d'Osaka), l'apport du téléport 1/1, permet *« de renouveler le genre du pavillon national, avec l'idée qu'il devait être présent sous différentes formes dans l'espace numérique »*. Et de rappeler que le spécialiste de la 3D n'est intervenu qu'une fois achevés les plans intérieurs et extérieurs du bâtiment par le cabinet d'architectes d'Olivier Coldefy, *« un an avant le démarrage du chantier, au moment de la finalisation de la scénographie, pour les derniers ajustements d'aménagement »*.

Ainsi, la plupart des fonctionnalités de la plateforme 3DEXPERIENCE étaient mobilisées pour ce défi, initié avec la filiale japonaise de Dassault Systèmes, partenaire Silver du Cofrex. Avant que le siège de Vélizy ne mette à sa disposition un téléport 1/1 dans lequel figurait un jumeau numérique du Pavillon France. Lequel fut modélisé avec les mêmes méthodes que pour la numérisation en 3D de la charpente de Notre-Dame-de-Paris avec ses 1,4 milliard de points à numériser. *« Le CNRS a bénéficié de la capacité à manipuler en immersion des vestiges des poutres calcinées numérisés en 3D juste après l'incendie. Les chercheurs du groupe de travail 'Bois' ont pu les manipuler, les déplacer virtuellement, y compris dans des zones très difficiles d'accès. Ce scan 3D réalisé par Andrew Tallon a permis de faire des comparaisons avant et après l'incendie. Il constituait la seule empreinte numérique tridimensionnelle existante de la cathédrale avant l'incendie »*, se souvient Medhi Tayoubi.

UN PAVILLON FRANCE « HYMNE À L'AMOUR »

Il doit incarner, jusqu'au 13 avril prochain, un hymne à l'amour sur 4 000 m². Situé à l'entrée d'un parc circulaire de 155 hectares clos par un gigantesque anneau de bois, le Pavillon France se veut, selon Emmanuel Macron, *« un hymne à l'audace et au dialogue »*, et doit servir d'écrin à l'excellence française. Cette construction éphémère, épurée et moderne avec ses façades latérales voilées, signée des équipes de Thomas Coldefy, sise sur Yumeshima Island à vingt minutes en métro du centre d'Osaka, abritera 1 000 m² d'exposition permanente, un bistrot, un espace VIP, etc. Le tout pour un budget de 53 millions d'euros, dont 13 millions sont apportés par des partenaires privés. L'ambition de ses promoteurs est d'y accueillir trois millions de visiteurs, physiques et virtuels.

Un outillage diversifié

Pour la maquette numérique et tridimensionnelle du Pavillon France, les scénographes, fournisseurs et sous-traitants du Cofrex bénéficiaient de Catia, Enovia, 3DVIA, Simulia, ces fonctionnalités tournant en C++ ou en Java, initiée avec SolidWorks, ASIS, Autodesk ou Abaqus. *« La solution 3D c'est révélé la plus intuitive pour que des équipes pluridisciplinaires puissent déambuler dans les espaces du pavillon, afin de se l'approprier, visualiser son aménagement et, au besoin, le modifier »*. Et ce, une fois capturées leurs déambulations via une multitude de caméras fixées aux murs.

Outre 3DEXPERIENCE, les acteurs du projet disposaient de casques de réalité virtuelle, pour visualiser en stéréoscopie les données 3D. Des outils collaboratifs comme des rayons laser permettaient aux participants de projeter des objets, eux aussi simulés en 3D, de les manipuler intuitivement. *« Ces casques ne possèdent que la puissance de calcul d'un smartphone. Le sac à dos l'accompagnant contient un ordinateur de gamer, le tout fonctionne avec des cartes graphiques Nvidia commandées en masse tout au long des séances de travail »*, rappelle Medhi Tayoubi.

Pour les utilisateurs du téléport, cette représentation grandeur nature des volumes du Pavillon France procurait nombre d'avantages. D'abord parce qu'une maquette numérique se révèle plus pratique et moins chère que sa jumelle physique, fragile et très compliquée à transporter. Ensuite, la 3D, surtout à l'échelle 1/1, démultiplie la visibilité. *« A l'échelle 1/100e, la hauteur de douze mètres du Pavillon France ne représente qu'une petite dizaine de centimètres. Il est donc difficile de se faire une idée de ce qu'elle représentera. Et encore plus des modifications nécessaires »*, insiste Jacques Maire. Surtout, elle offrait la possibilité de simuler avec le plus grand réalisme les flux des visiteurs ou les modifications suggérées. *« Nous nous sommes rendu compte que dans l'espace VIP, les accès aux issues de secours depuis les toilettes étaient trop justes »*, pointe le patron du Cofrex. Qui se souvient aussi que cela a permis de rehausser la hauteur d'un bassin dans le jardin, pour améliorer le confort des visiteurs. Avant d'avouer : *« nous aurions aimé profiter du téléport dès la conception du bâtiment »*.

Le téléport devait considérablement fluidifier les relations entre les parties prenantes au projet. Il s'est rapidement imposé comme un vecteur de communication auprès de futurs partenaires. Restait à tempérer l'enthousiasme des uns et des autres : *« le plus gros défi fut de réduire le temps de présentation du pavillon à vingt minutes »*, sourit Jacques Maire.

Une vue du bâtiment

Une image de marque

Alors que le chantier 'réel' du Pavillon France s'achève à Osaka, son jumeau virtuel n'est pas près d'être remisé. Tokyo accueillera un troisième téléport, après Vélizy et l'Avenue de La Grande Armée, pour projeter le savoir-faire de Dasty dans l'imagerie médicale en 3D immersive. Une solution de rendu déporté par wifi, effectuant le calcul à distance et sans latence, ne nécessitera que le port du casque. Dans le même temps, des visites 'virtuelles' seront proposées partout dans le monde, même depuis un smartphone. Surtout, la Cité de l'Architecture et du Patrimoine archivera ce 'nuage de points' disponible pour montrer aux étudiants et chercheurs, *« le Pavillon France tel qu'il était visitable et tel qu'il aurait dû être, et qui sait, inspirer de futurs pavillons pour de futures expositions universelles »*, suggère Jacques Maire.

Source de projets

L'expérience acquise avec le Pavillon France irrigue deux projets en cours de Dasty. D'abord, pour l'extension de l'Hôpital américain de Neuilly, en collaboration avec les équipes de Jean-Michel Willemotte. *« Le téléport s'est révélé un outil primordial pour la conduite du changement, pour embarquer les équipes de soignants de ce bâtiment avant même qu'il ne soit terminé : en scannant les blocs opératoires existants et en projetant les utilisateurs dans ces futurs 'plateaux' chirurgicaux, pour faciliter le dialogue et la prise de décision avant les commandes de matériel »*, raconte Medhi Tayoubi. Dans un autre registre, 3DEXPERIENCE a renoué avec le savoir-faire originel de sa maison-mère. Le fabricant de cosmétiques L'Occitane a en effet sollicité Dassault Systèmes pour 'projeter' une nouvelle ligne d'assemblage sur un site de production. Les simulations ont mis en évidence treize points de sécurité jusque-là omis ou restés invisibles dans des solutions 2D.☐

V.B

GUIDELINE

2017 : Mise au point du téléport. Partenariat avec La Cité de l'Architecture et du Patrimoine

Été 2024 : Lancement du téléport pour L'Exposition universelle d'Osaka

4 février 2025 : Conférence de présentation du Pavillon France et de la délégation française à L'Exposition universelle d'Osaka, en présence d'Emmanuel Macron à La Cité de l'Architecture et du Patrimoine, démonstration du téléport

1er février 2025 : Démarrage du chantier du Pavillon France

13 avril — 13 octobre 2025 : Exposition universelle d'Osaka

Quinzaine médicale

ML Qu'est-ce que la régression linéaire et comment est-elle utilisée ?

Le machine learning s'appuie sur différents concepts mathématiques. Parmi ceux-ci, l'un des plus simples mais néanmoins efficaces pour certaines phases est la régression linéaire. Nous allons voir, dans cet article, quels en sont les principes, dans quel cas et de quelle façon elle est utilisée en ML.

Plusieurs techniques issues de la statistique et de la probabilité ont permis d'accroître les connaissances sur l'analyse de données, la suppression de données aberrantes ou la manière de gérer des données manquantes pour choisir une représentation pertinente d'un phénomène.

Le machine learning

Une fois les données bien préparées se pose la question de comment extraire des informations intéressantes et efficaces sur des grands volumes de données. Cela nécessiterait normalement des procédures très gourmandes en ressources informatiques. C'est là qu'intervient le machine learning, ou apprentissage automatique, qui permet de rendre un programme capable d'apprendre à partir d'exemples de données, sans être programmé pour cela initialement. Cette définition originelle du machine learning nous vient d'Arthur Samuel (1959). L'idée est relativement facile à comprendre : elle est similaire à l'apprentissage d'un être humain qui, d'expérimentations en expérimentations successives sur un même sujet, apprend de lui-même progressivement et devient de plus en plus compétent et efficace. Mais ce qui coule de source pour un humain est moins évident pour une machine. En pratique, faire apprendre à une machine ou un robot à apprendre lui-même nécessite des connaissances en mathématiques, plus précisément en statistique et en probabilité, pour construire des modèles, et en informatique pour implémenter des algorithmes efficaces et robustes d'auto-apprentissage. Une fois l'algorithme implémenté, la machine va pouvoir apprendre et prédire des phénomènes précis et s'enrichir au fur et à mesure qu'elle reçoit de nouvelles données. C'est ce que l'on va alors qualifier de « système intelligent ». Il faut bien entendu être prudent avec le terme intelligent et ne pas l'entendre comme l'intelligence humaine. Pas à ce stade, en tous cas. Le machine learning, communément appelé ML, est devenu l'une des branches principales de l'intelligence artificielle, simplement parce qu'il permet d'obtenir des résultats concrets, alors que d'autres branches plus complexes n'ont pas encore atteint ce stade de production. Le ML peut être divisé en plusieurs catégories ou types de problématiques : l'apprentissage supervisé, l'apprentissage semi-supervisé et l'apprentissage non supervisé. Il englobe notamment un autre concept lui aussi très à la mode : le deep learning, ou apprentissage

LES CONCEPTS MATHÉMATIQUES DERRIÈRE LE MACHINE LEARNING

LA RÉGRESSION LINÉAIRE

Un modèle de régression linéaire simple est de la forme :

$$Y = aX + b + \varepsilon \text{ où } f(X) = aX + b$$

Avec :

- Y, la variable cible, aléatoire dépendante
- a et b, les coefficients (pente et ordonnée à l'origine) à estimer
- X, la variable explicative, indépendante
- ε, une variable aléatoire qui représente l'erreur

Un modèle de régression linéaire multiple est de la forme :

$$Y = ax_1 + bx_2 + cx_3 + \cdots + K + \varepsilon \text{ où } f(X) = aX + b$$

Avec :

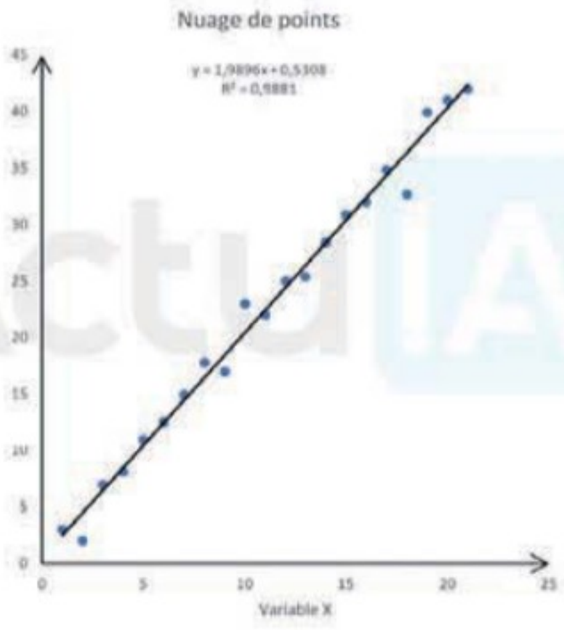
- Y, la variable cible, aléatoire dépendante
- $a,\ldots,K$, les coefficients (pente et ordonnée à l'origine) à estimer
- $X = (x_1, \ldots, x_q)$, la variable explicative, indépendante
- εune variable aléatoire qui représente l'erreur

Sous forme matricielle, le modèle de régression linéaire simple est de la forme

$$Y = AX + \varepsilon$$

Où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_p \end{pmatrix}, A = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$



Le terme R^2 de l'image représente le coefficient de corrélation de Bravais-Pearson au carré. Ce coefficient mesure l'intensité de la relation.

$$R = \frac{\sum_{i=1}^{p}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{p}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{p}(Y_i - \bar{Y})^2}}$$

Le principe des moindres carrés ordinaires consiste à choisir les valeurs de a et b qui minimisent les erreurs de prédiction ou les résidus sur un jeu de données d'apprentissage :

$$\varepsilon = \sum_{i=0}^{p}\left(Y_i - (aX_i + b)\right)^2$$

Minimiser cette expression revient à résoudre un problème d'optimisation, voici la forme des estimateurs notés $\hat{a}$ et $\hat{b}$

$$\hat{a} = \frac{\sum_{i=1}^{p}(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2} = \frac{c_{xy}}{s_x^2}$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}$$

Où c_{xy} est la covariance empirique entre les X_i et les Y_i et s_x^2 est la variance empirique des X_i.

L'expression de $\hat{b}$ indique que la droite de régression linéaire passe par le centre de gravité du nuage de point $(\bar{X}, \bar{Y})$.

Source: www.actuia.com

Vous trouverez des informations intéressantes sur la régression linéaire sur le site d'actuia.com comme cette fiche pratique à l'adresse https://www.actuia.com/storage/uploads/2019/06/la-regression-lineaire.png

profond. La régression linéaire est l'un des concepts de base du machine learning, mais pas le seul.

Qu'est-ce que la régression linéaire ?

La régression linéaire est une technique d'analyse de données capable de prédire la valeur de données inconnues, en utilisant un ensemble de valeurs de données apparentées et connues. Elle modélise mathématiquement la variable inconnue ou dépendante et la variable connue ou indépendante sous la forme d'une équation linéaire. Si, par exemple, vous disposez de données sur vos dépenses et vos revenus de l'année dernière, les techniques de régression linéaire pourront analyser ces données et déterminer que vos dépenses représentent la moitié de vos revenus. Elles calculeront ensuite une dépense future inconnue en réduisant de moitié un revenu futur connu. C'est bien entendu un cas extrêmement simple, mais c'est le principe de base. Les modèles de régression linéaire sont relativement simples et fournissent une formule mathématique facile à interpréter pour générer des prévisions. La régression linéaire est une technique statistique bien établie qui s'applique aisément aux logiciels et au calcul. Les entreprises y ont recours pour convertir de manière fiable et prévisible les données brutes en informatique décisionnelle en informations exploitables. Les scientifiques de nombreux domaines, biologie, sciences comportementales, environnementales, sociales et autres utilisent la régression linéaire au moins pour effectuer une analyse préliminaire des données connues et prédire des tendances futures. De nombreuses méthodes de science des données, l'intelligence artificielle en général et le machine learning en particulier, utilisent la régression linéaire pour résoudre des problèmes complexes.

Pour en savoir encore plus sur la régression linéaire et télécharger des données d'exemples pour travailler avec, rendez-vous sur le site de datatab à l'adresse https://datatab.fr/tutorial/linear-regression

Comment fonctionne la régression linéaire ?

Fondamentalement, une technique de régression linéaire simple tente de tracer un graphique linéaire entre deux variables de données, X et Y. En tant que variable indépendante, X est tracée le long de l'axe horizontal. Les variables indépendantes sont également appelées variables explicatives ou variables prédictives. La variable dépendante, Y, est tracée sur l'axe vertical. Vous pouvez également faire référence aux valeurs Y en tant que variables de réponse ou variables prédites.

Relation linéaire

Une relation linéaire doit exister entre les variables indépendantes et dépendantes. Pour déterminer cette relation, les scientifiques des données créent un nuage de points, une collection aléatoire de valeurs X et Y, pour voir s'ils se situent le long d'une ligne droite. Si tel n'est pas le cas, vous pouvez appliquer des fonctions non linéaires telles que la racine carrée ou le journal pour créer mathématiquement la relation linéaire entre les deux variables.

L'algorithme de régression linéaire

L'algorithme de régression linéaire est un algorithme d'apprentissage supervisé. Cela signifie qu'à partir de la variable cible ou de la variable à expliquer nommée Y, le modèle a pour but de faire une prédiction grâce à des variables dites explicatives X ou prédictives. Concrètement, cela va conduire à prédire la valeur de vente d'une maison en fonction de sa superficie, de sa localisation, de la présence ou non d'un parking, ou de prédire le potentiel de croissance d'une plante en fonction de l'ensoleillement, du vent, de la composition de la terre, de l'arrosage, de la présence d'autres plantes autour, de celle de parasites sont deux exemples parmi une infinité d'utilisation du modèle de régression linéaire. Voyons un peu, en termes simples, à quoi ressemble le modèle de régression linéaire. Un modèle de régression linéaire est un modèle de machine learning dont la variable cible Y est quantitative, tandis que la variable X peut, elle, être quantitative ou qualitative. L'objectif va être de trouver une fonction dite de prédiction ou une fonction coût qui décrit la relation entre X et Y, c'est-à-dire qu'à partir de valeurs connues de X, on arrive à donner une prédiction des valeurs de Y. La fonction recherchée sera de la forme : Y= f(X), f(X) étant une fonction linéaire.

À partir d'un échantillon de points qui vont représenter les données connues, il va falloir répartir les données en deux groupes : les données d'entraînement et les données de test. La première catégorie de données (celles d'entraînement) servira pendant la phase d'apprentissage du modèle

alors que la seconde (celles de test) sera utilisée pour évaluer la qualité de prédiction du modèle. Le but n'est donc pas de construire une fonction qui prédira avec une précision optimale les valeurs des variables cibles, mais une fonction qui se généralisera au mieux pour prédire des valeurs de données qui n'ont pas encore été observées. Avant de débuter une étude de régression simple, il faudra d'abord tracer les observations de type (Xi, Yi), i=1..., p et c'est ce que va faire l'algorithme de régression linaire.

L'avantage de l'algorithme de régression linéaire est sa grande simplicité d'interprétation et sa facilité de calcul. Le data scientist devra en revanche bien vérifier qu'il existe une relation linéaire entre les paramètres d'entrée et celui de sortie. Le modèle présente quelques inconvénients, comme le fait que l'algorithme soit très sensible aux valeurs aberrantes (outliers) des données d'apprentissage, d'où la nécessité de bien préparer ses données dès le départ. Il existe des méthodes dites de régularisation permettant de pallier ce problème. Les méthodes de régularisation servent à pénaliser les valeurs trop grandes des coefficients ai et b. En plus de cela, le caractère linéaire du modèle néglige les interactions entre les variables explicatives. C'est pour cela qu'il existe une possibilité de définir de nouvelles variables explicatives comme étant le produit de variables existantes.

Modèle de la régression linéaire

Modélisation	Nature de la régression
Une seule variable explicative X	Régression simple
Plusieurs variables explicatives Xj (j=1,..., q)	Régression multiples

Le modèle de régression linéaire analyse les relations entre la variable dépendante ou variable cible Y et l'ensemble des variables indépendantes ou explicatives X. Cette relation est exprimée comme une équation qui prédit les valeurs de la variable cible comme une combinaison linéaire de paramètres.

Un modèle de régression linéaire simple est de la forme :
$Y=aX+b+\varepsilon$ où $f(X)= aX+b$

Avec :
Y, la variable cible, aléatoire dépendante ;
a et b, les coefficients (pente et ordonnée à l'origine) à estimer ;
X, la variable explicative, indépendante ;
ε, une variable aléatoire qui représente l'erreur.

Un modèle de régression linéaire multiple est de la forme :
$Y=ax_1+bx_2+cx_3+\bullet\bullet\bullet+K+\varepsilon$ où $f(X)= aX+b$

Avec :
Y, la variable cible, aléatoire dépendante ;
a,..., K les coefficients (pente et ordonnée à l'origine) à estimer ;
$X=(x_1,..., x_q)$, la variable explicative, indépendante ;
ε, une variable aléatoire qui représente l'erreur.

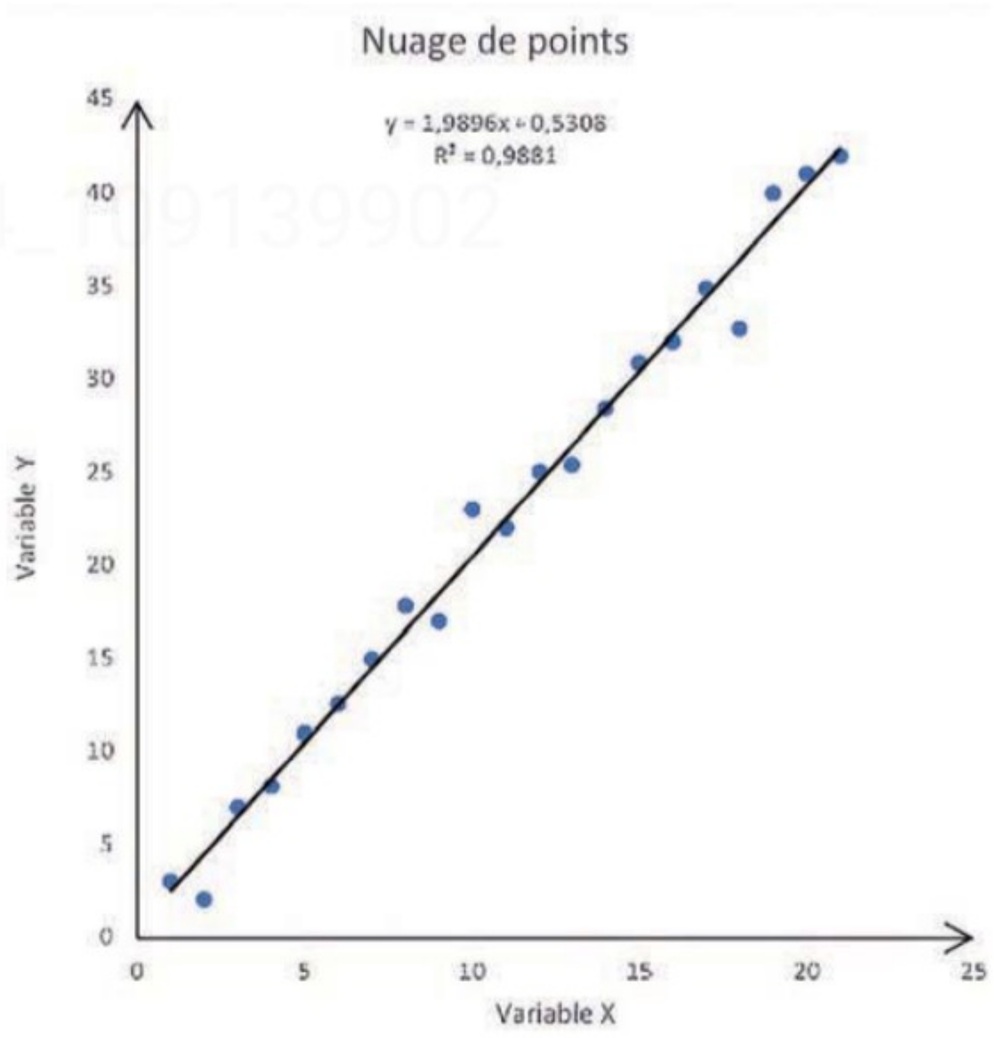
Sous forme matricielle, le modèle de régression linéaire simple est de la forme :
$Y=AX+\varepsilon$

Où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ . \\ y_p \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ & \vdots \\ & . \\ 1 & x_p \end{pmatrix}, A = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ . \\ \varepsilon_p \end{pmatrix}$$

Avec :
Y, un vecteur à expliquer de taille n x 1,
X, la matrice explicative de taille n x 2,
ε, le vecteur d'erreurs de taille n x 1

ε est appelé résidu. C'est l'erreur commise, c'est-à-dire l'écart entre la valeur Yi observée et la valeur a_iX_i+b donnée par la relation linéaire. En effet, même si une relation linéaire est effectivement présente, les données mesurées ne vérifient en général pas exactement cette relation. Pour ce faire, on tient compte des erreurs observées dans le modèle mathématique.



Sur ce graphique, la droite de régression linéaire ou la droite des moindres carrés de Y en X représente la droite d'ajustement linéaire, celle qui résume le mieux la structure du nuage de points pendant la phase d'apprentissage. Elle rend minimale la somme des carrés des erreurs d'ajustement.

C'est en confrontant l'équation calculée par l'algorithme de régression linéaire aux nouvelles données de la réalité (X), que les prédictions (Y) seront réalisées par l'algorithme d'intelligence artificielle en production.

Le terme R^2 de l'image représente le coefficient de corrélation de Bravais-Pearson au carré. Ce coefficient mesure l'intensité de la relation linéaire entre Y et X.

Voici sa formule :

$$R = \frac{\sum_{i=1}^{p}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{p}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{p}(Y_i - \bar{Y})^2}}$$

Le coefficient de corrélation est un nombre toujours compris entre -1 et 1.
Si R est proche de 1 : il y a une forte liaison linéaire entre les variables, et les valeurs prises par Y ont tendance à croître quand les valeurs de X augmentent.
Si R est proche de 0 : il n'y a pas de liaison linéaire,
Si R est proche de -1 : il y a une forte liaison linéaire et les valeurs prises par Y ont tendance à décroître quand les valeurs de X augmentent.

Le coefficient de corrélation mesure la qualité de la droite d'ajustement linéaire mais ne représente en aucun cas une cause de la relation logique entre X et Y. Seul le data scientist sera à même d'estimer la relation logique entre les deux variables. Un exemple souvent cité est : la mortalité des jeunes enfants désespère les femmes qui renoncent à lire Freud... Même si le coefficient de corrélation est supérieur à 0, il n'y a a priori aucun lien logique entre les deux phénomènes. On trouve fréquemment d'ailleurs, sur des sites internet, des corrélations farfelues de ce genre : il ne faut pas, à la base, comparer n'importe quoi avec n'importe quoi, et là, le rôle de l'humain est essentiel. Il faut donc bien faire la distinction entre corrélation et causalité.

Indépendance résiduelle

Les scientifiques des données utilisent des valeurs résiduelles pour mesurer la précision des prévisions. Une valeur résiduelle est la différence entre les données observées et la valeur prédite. Les valeurs résiduelles ne doivent pas présenter de schéma identifiable entre elles. Par exemple, vous ne voulez pas que les valeurs résiduelles augmentent avec le temps. Vous pouvez utiliser différents tests mathématiques, tels que le test de Durbin-Watson, pour déterminer l'indépendance résiduelle. Vous pouvez utiliser des données factices pour remplacer toute variation de données, telle que les données saisonnières.

Normalité

Les techniques graphiques telles que les diagrammes Q-Q déterminent si les valeurs résiduelles sont normalement distribuées. Les valeurs résiduelles doivent se situer le long d'une ligne diagonale au centre du graphique. Si les valeurs résiduelles ne sont pas normalisées, vous pouvez tester les données pour rechercher des valeurs aberrantes aléatoires ou des valeurs qui ne sont pas typiques. La suppression des valeurs aberrantes ou l'exécution de transformations non linéaires peuvent résoudre le problème.

ESTIMATION DES COEFFICIENTS DE LA DROITE : MÉTHODE DES MOINDRES CARRÉS

La régression linéaire est relativement simple d'un point de vue mathématique. Ce qui fait que ce type d'algorithme entre pleinement dans le cadre du machine learning est le fait qu'un logiciel soit capable d'ajuster les paramètres a et b à partir d'exemples fournis par l'utilisateur. Mais voyons comment ces paramètres sont ajustés afin d'estimer la variable de sortie Y :

$$\varepsilon = \sum_{i=0}^{P}\left(Y_i - (aX_i + b)\right)^2$$

Le principe des moindres carrés ordinaires consiste à choisir les valeurs de a et b qui minimisent les erreurs de prédiction (ou résidus) sur un jeu de données d'apprentissage :
Minimiser cette expression revient à résoudre un problème d'optimisation, voici la forme des estimateurs notés â et $\hat{b}$ qui sont égaux à :

$$\hat{a} = \frac{\sum_{i=1}^{p}(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2} = \frac{c_{xy}}{s_x^2}$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}$$

Où c_{xy} est la covariance empirique entre les X_i et les Y_i, et S_x^2 est la variance empirique des X_i. L'expression de $\hat{b}$ indique que la droite de régression linéaire passe par le centre de gravité du nuage de points ($\bar{X}$, $\bar{Y}$).

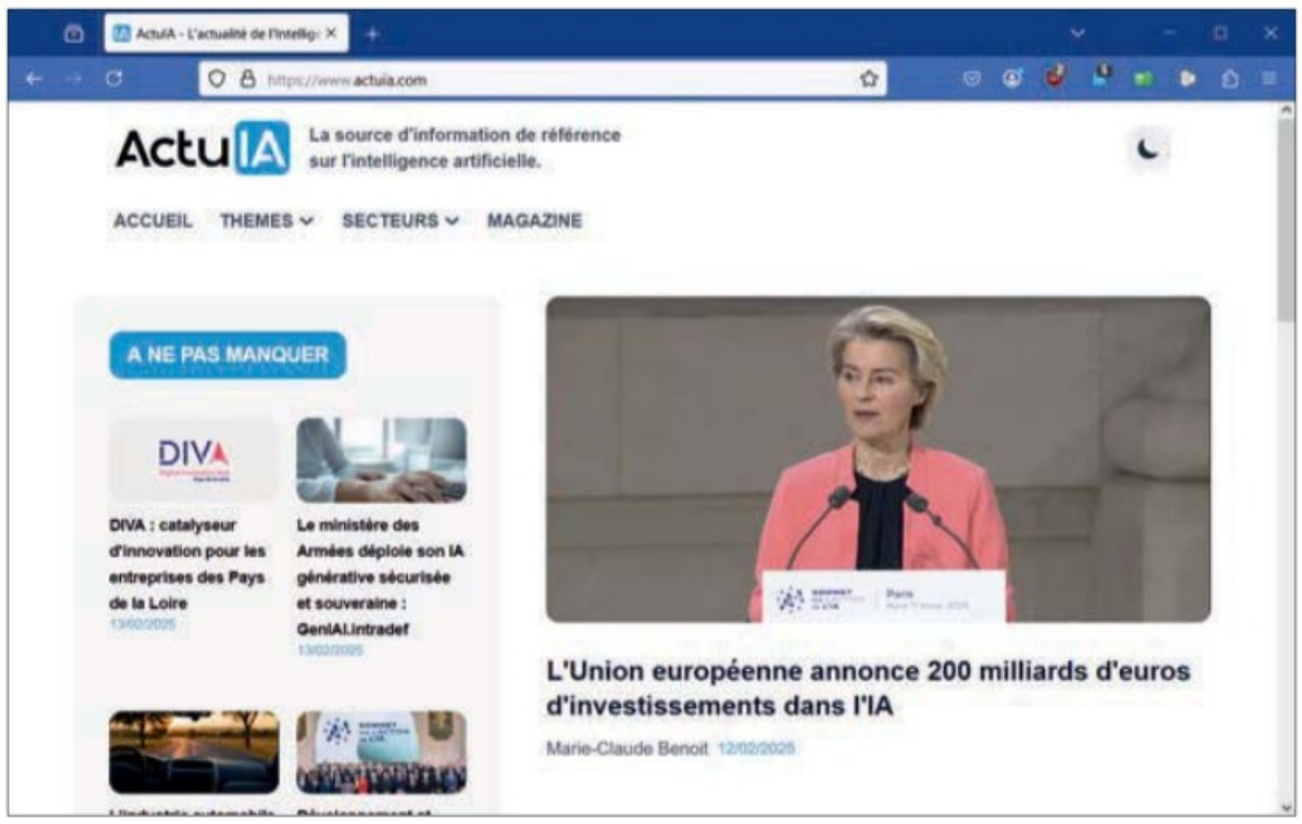
Homoscédasticité

L'homoscédasticité suppose que les valeurs résiduelles ont une variance constante ou un écart type par rapport à la moyenne pour chaque valeur de x. Dans le cas contraire, les résultats de l'analyse risquent de ne pas être exacts. Si cette hypothèse n'est pas respectée, vous devrez peut-être modifier la variable dépendante. Comme la variance se produit naturellement dans les grands jeux de données, il est logique de modifier l'échelle de la variable dépendante. Par exemple, au lieu d'utiliser la taille de la population pour prédire le nombre de casernes de pompiers dans une ville, vous pouvez utiliser la taille de la population pour prédire le nombre de casernes de pompiers par personne. Certains types d'analyse de régression sont plus adaptés que d'autres pour gérer des jeux de données complexes. En voici quelques exemples ci-dessous.

Régression linéaire simple

La régression linéaire simple est définie par la fonction linéaire : Y= β0*X + β1 + ε

β0 et β1 sont deux constantes inconnues représentant la pente de régression, tandis que ε (epsilon) est le terme d'erreur. Vous pouvez utiliser une régression linéaire simple pour modéliser la relation entre deux variables, telles que celles-ci :
Précipitations et rendement des cultures ;
Âge et taille des enfants ;
Température et expansion du mercure métallique dans un thermomètre.



Le site d'actuia.com (https://www.actuia.com/) est une mine d'informations sur l'IA

Régression linéaire multiple

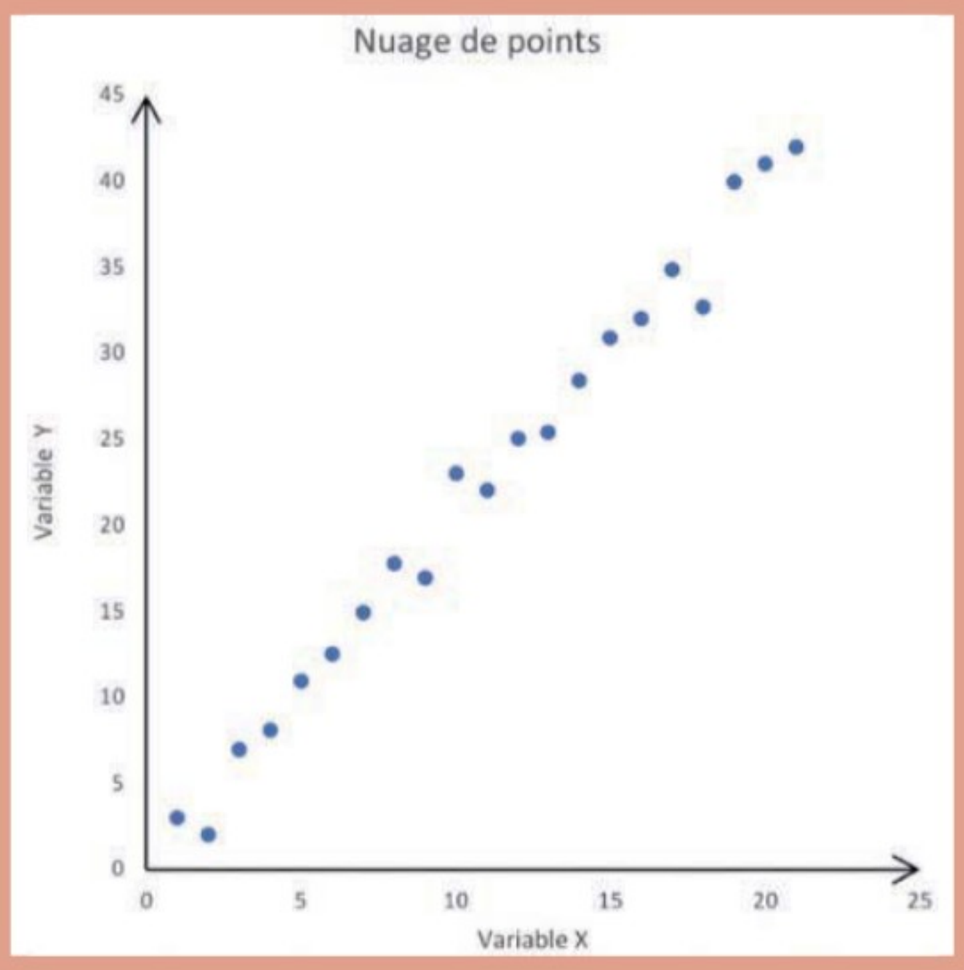
Dans l'analyse de régression linéaire multiple, le jeu de données contient une variable dépendante et plusieurs variables indépendantes. La fonction de la droite de régression linéaire change pour inclure davantage de facteurs, comme ci-dessous :
Y= β0*X0 + β1X1 + β2X2+..... βnXn+ ε

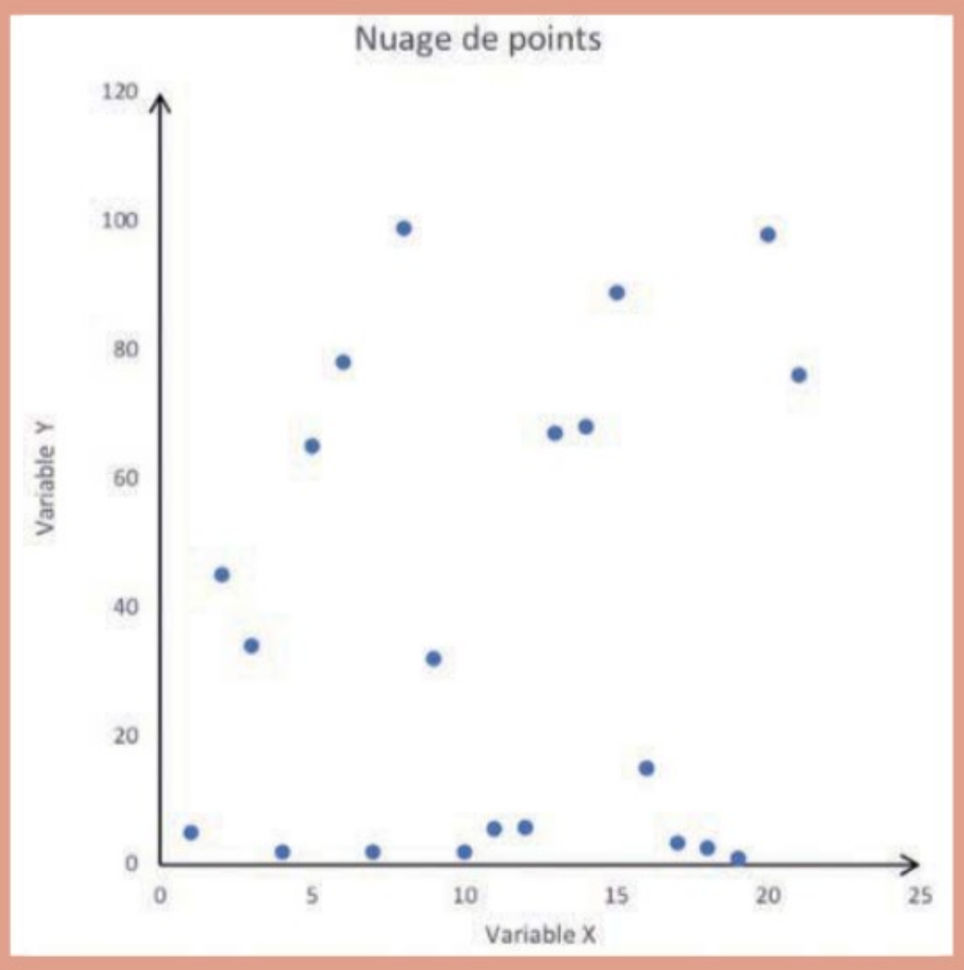
À mesure que le nombre de variables prédictives augmente, les constantes β augmentent également en conséquence.

La régression linéaire multiple modélise plusieurs variables et leur impact sur un résultat :
Utilisation des précipitations, de la température, de l'ensoleillement, de la richesse du sol et des engrais sur le rendement des cultures ;
Influence du régime alimentaire et de l'exercice sur les maladies cardio-vasculaires ;
Influence de la croissance des salaires et de l'inflation des prix des carburants sur l'inflation en général. □ **T.T**

REPRÉSENTATION GRAPHIQUE

Le but est de savoir si le modèle linéaire est oui ou non pertinent pour l'étude d'un phénomène donné. Le graphique est au départ un nuage de points sur lequel on va tenter de relever la tendance qu'a sa forme.





En regardant ces deux graphiques, il semble clairement approprié d'utiliser le modèle linéaire pour la première image. Pour la deuxième, en revanche, aucune tendance connue ne ressort. Ce n'est pas toujours aussi évident, bien entendu, et il y aura des cas par exemple où la régression linéaire pourra être employée mais ne sera pas suffisante.

Recrutement

Les employeurs veulent du savoir-être chez les jeunes ingénieurs

Les recruteurs souhaitent embaucher de jeunes diplômés d'écoles d'ingénieurs en informatique avec un bon niveau technologique, mais aussi autonomes et adaptables.

« *Les attentes des employeurs sont aussi fortes sur les soft skills que sur les compétences techniques, dans un monde en évolution rapide, plein d'incertitudes. Ils souhaitent des jeunes diplômés autonomes et adaptables, capables de travailler en équipe, si ce n'est d'animer une équipe* », explique Sylvain Goussot, directeur général de l'école d'ingénieurs Epita, analysant une enquête IPSOS qu'il a commanditée. Publiée en début d'année, elle porte sur la perception des diplômés d'écoles d'ingénieurs en informatique par 301 d'entreprises de 100 salariés et plus responsables (à la DG, aux RH ou à la DSI). Une moitié est dans les activités informatiques, de conseil et d'ingénierie, l'autre moitié est tous secteurs.

Savoir-être et expertise technique exigés

Les soft skills que doivent montrer en priorité les jeunes diplômés, d'après respectivement les dirigeants tous secteurs et ceux des sociétés informatiques sont l'adaptabilité (50/51 %), l'autonomie (54/44 %), la rigueur (42/44 %) et la curiosité (39/30 %). Puis viennent la recherche de solutions (28/39 %), l'implication (29/36 %), l'aisance relationnelle (28/23 %), et la créativité (20/23 %). Les attentes sont donc loin de l'image de l'informaticien geek asocial.

Ce savoir-être affirmé doit accompagner, sans surprise, un haut niveau d'expertises techniques. Les compétences prioritaires sont : maîtriser les environnements de développement et de production informatique (58 % pour les entreprises de tous secteurs/47 % chez les sociétés d'ingénierie informatique), piloter et gérer des projets collaboratifs (45/48 %), étudier, analyser et concevoir des solutions robustes (44/45 %), gérer les risques dans les systèmes complexes (45/42 %). Viennent ensuite le travail en environnement agile sur des problématiques d'innovation (38/39 %), et l'intégration des enjeux de Green IT (22/30 %).

En pratique, les responsables des entreprises de tous secteurs jugent plus sévèrement que leurs homologues de l'ingénierie informatique les réponses des jeunes ingénieurs à leurs attentes : les premiers sont plus nombreux à estimer qu'ils répondent assez mal ou très mal à leurs attentes sur certaines compétences techniques et soft skills (même si cela représente une minorité des répondants), les seconds sont plus nombreux à dire qu'ils répondent très bien aux attentes (cf. graphiques). « *Les attentes des ESN sont globalement un peu moins fortes,* commente S. Goussot, *car elles ont l'habitude d'embaucher de nombreux jeunes qu'elles forment, tandis que les entreprises veulent des personnes opérationnelles très rapidement, 'prêtes à l'emploi'.* »

Sylvain Goussot, directeur général de l'Epita

Des enjeux clefs à maîtriser

Les entreprises ont été également interrogées sur les principaux enjeux des DSI aujourd'hui. L'enjeu clef est la protection contre les cyberattaques (84 % pour les sociétés de tous secteurs/80 % parmi les entreprises d'ingénierie informatique), loin devant la digitalisation et l'automatisation des processus métiers (54/59 %), et les transformations technologiques à l'échelle de l'entreprise (50/52 %). Représentent moins de défis opérer une infrastructure cloud pérenne (32/24 %), déployer l'intelligence artificielle à l'échelle de l'entreprise (24/32 %) et adopter les pratiques IT durables (27/20 %).

« *Les résultats sont décevants sur le green IT et l'IA,* regrette M. Goussot, *nous espérons que ces enjeux remonteront dans l'agenda des DSI. Nous prenons de toute façon en compte les besoins des entreprises, en les écoutant en rendez-vous, et en leur donnant la parole sur nos programmes au sein des conseils de perfectionnement. Nous ouvrons ainsi une spécialisation sur le cloud souverain au sein du cursus ingénieur en février 2026.* » ☐

C.C

« Nous recrutons beaucoup de développeurs et de Tech leads »

Rencontre avec Sylvie Verstraeten, DRH Docaposte

La filiale de solutions numériques de confiance du groupe La Poste embauche près de 500 profils technologiques cette année en CDI, en particulier en développement, cybersécurité et en data.

L'Informaticien : Pouvez-vous présenter votre entreprise ?
Sylvie Verstraeten : Docaposte, filiale du groupe La Poste, a réalisé un chiffre d'affaires de 879 millions d'euros en 2024, comptant 60 000 clients entreprises et administrations. Elle conçoit, développe et opère des solutions numériques de confiance et de souveraineté, et offre des services et du conseil. Elle compte plus de 6 000 collaborateurs dans 18 pays en Europe, en Afrique du Nord et en Amérique. La majeure partie travaille sur 95 sites en France. Nos effectifs se répartissent à 30 % dans l'édition de logiciels, 28 % dans la partie ESN (dont conseils et services), 25 % dans l'intégration et le back office numérique. Nous comptons environ 1 800 ingénieurs IT, 1 000 développeurs dans différents langages, 400 experts métiers, 400 experts en data et intelligence artificielle, et 80 spécialistes en UX/UI. Notre turnover est de 12 %, tous postes confondus.

Quelle est votre politique de recrutement ?
Après plusieurs années sur un marché du travail particulièrement tendu où nous avons recruté près de 1000 profils par an, depuis 2024, nous recrutons 600 personnes en CDI et plus de 150 alternants et stagiaires par an (60 % sont ensuite recrutés en CDI), dont près de 80 % dans les technologies. Les profils les plus recherchés sont les développeurs et Tech leads, devant les ingénieurs cybersécurité et data scientists. Notre 9e édition de Master Dev France, qui valorise les métiers du développement à travers des conférences et un concours de code, a réuni plus de 800 développeurs en mars.

Les jeunes diplômés comptent pour 20 % de nos recrutements en CDI. Nous sommes partenaires d'écoles d'ingénieurs et d'informatique : Epitech, Efrei, le pôle Léonard de Vinci, les Gobelins, Simplon et Ada Tech School.

Nous recrutons à tous niveaux d'expérience. Comme nos confrères, nous avons des difficultés de recrutement en développement, cybersécurité et data science.

Comment se déroule le recrutement ?
Notre processus de recrutement est classique et fluide. Il va de 30 jours pour un profil junior, à 60 jours en moyenne tous postes confondus. Nous soignons nos offres d'emploi, détaillées et transparentes. Notre source principale de profils expérimentés est LinkedIn. Après un premier entretien avec l'un de nos recruteurs internes, le candidat passe un

© David Arous

ou deux entretiens avec des managers opérationnels et RH. Nous passons par des cabinets de recrutement pour trouver certains profils seniors. 95 % des candidats se disent satisfaits de nos procédures de recrutement.

Quelles sont les attentes des candidats ?
Leur principale motivation est l'intérêt du poste, suivi par un environnement de travail stimulant et agréable. Ils sont intéressés par nos projets à impact, nos projets de numérique de confiance, par exemple dans le vote électronique. Nous sommes transparents sur nos conditions. Ainsi, sur des projets à forts enjeux, nous insistons sur l'investissement élevé à fournir.
Les candidats, en particulier les jeunes, veulent de l'autonomie, du sens, de l'innovation et la flexibilité du travail. Nos accords d'entreprise permettent de télétravailler deux à trois jours par semaine selon les postes. Et nous ouvrons nos grands sites de Paris La Défense et de Sophia Antipolis à ceux ayant besoin de travailler en dehors de leur site de rattachement.

Comment se passe l'intégration ?
Le premier jour, la recrue est accueillie et accompagnée par son manager et son responsable RH. Elle se rend ensuite à une matinée d'intégration à notre siège

d'Ivry-sur-Seine, lui permettant de créer un premier réseau. Elle dispose sur notre Intranet de vidéos d'informations et de formations obligatoires (éthique des affaires, RSE, bien-être au travail...), suivies généralement d'un quiz. La période d'essai est de quatre mois renouvelable, pendant laquelle elle a des entretiens managériaux et RH.

Quelle est votre politique de fidélisation ?

Elle s'articule autour de plusieurs piliers. Dans notre baromètre d'engagement, auquel a répondu 76 % de nos collaborateurs, la qualité de vie au travail est noté 7,2 sur 10. 84 % déclarent avoir un environnement de travail motivant, et 89 % de bonnes relations avec leur manager. Nos espaces de travail ont été pensés pour faciliter la collaboration.

La proximité est clef. L'accès à l'information et au management est facilité. Le comité exécutif élargi, qui comprend une centaine de managers, se réunit lors de séminaires. Les managers de tous niveaux, soit environ 1 000 personnes, participent à notre convention annuelle. Le comité exécutif va à la rencontre des collaborateurs sur les sites. Notre Pdg, Olivier Vallet, et moi-même visitons les sites et organisons des cafés informels pour discuter avec 12 personnes au maximum.

Nous proposons aussi des moments d'échange : « lundi digital » sur un sujet digital (comment faire son personal branding sur LinkedIn, comment rédiger un bon prompt...), parfois animé par un intervenant externe, « mardi solution » sur la stratégie ou l'une de nos solutions.

Equipes et sites organisent des événements conviviaux, par exemple à Sophia Antipolis autour de la galette des rois et du carnaval de Nice.

L'innovation est également essentielle. Des comités d'innovation ont lieu un jeudi par mois, où tous les collaborateurs peuvent participer.

UNE ÉCOLE DATA ET IA À LA POSTE DEPUIS 2023

L'Ecole de la data et de l'IA du groupe La Poste a intégré depuis 2023 une centaine d'apprenants engagés dans un cursus certifiant. 50 collaborateurs ont suivi ou suivent un parcours de reconversion interne sur une fonction data (data analyst, data engineer, data scientist). Et 50 alternants ont bénéficié d'un contrat d'apprentissage ou de professionnalisation en lien avec un cursus sur les métiers de la data et de l'IA. Dans un principe de parité, l'école accueille pour moitié des femmes.

Nous avons, enfin, une politique de mobilité forte. Nous ouvrons nos postes en interne avant de les proposer en externe. Et en 2024, nous avons accompagné 150 collaborateurs qui ont souhaité évoluer vers un autre domaine ou métier.

Quelle est votre politique d'inclusion et de diversité ?

Le premier de nos trois axes est l'égalité professionnelle et la parité femmes-hommes. Docaposte comprend 41% de femmes, 31% dans les métiers technologiques et 30 % au comité exécutif. Notre score à l'index de l'égalité professionnelle est de 86 sur 100. Nous sommes vigilants sur les recrutements, promotions et rémunérations. Nous avons lancé un concours de cooptation où nous doublons les primes pour attirer plus de femmes vers les métiers technologiques. Nous avons travaillé lors d'un hackathon sur le sujet de la promotion des femmes aux postes de direction.

Sur le handicap, nous réalisons différentes actions de sensibilisation, via le gaming et les jeux. Lors du Duoday, des personnes en situation de handicap découvrent un métier pendant une journée, accompagnées chacune par un collaborateur. Si nos effectifs comptent un peu moins de 4 % de personnes en situation de handicap, nous compensons en faisant appel à des ESAT.

Nous sommes signataires de la Charte de la diversité.

Enfin, nous avons un engagement sociétal, contribuant notamment à l'insertion des jeunes des zones « oubliées ». Notre programme *« Incarner le numérique responsable »* est en faveur d'un numérique éthique, responsable et inclusif. Docaposte soutient ainsi à partir du printemps 2025 Emmaüs Connect dans l'élaboration d'une méthodologie d'évaluation de l'inclusivité des services publics numériques, puis dans son expérimentation avec deux groupes de personnes en situation d'exclusion numérique et sociale, et enfin dans la production d'un guide méthodologique et de fiches ressources à destination des services publics et acteurs sociaux. □

C.C

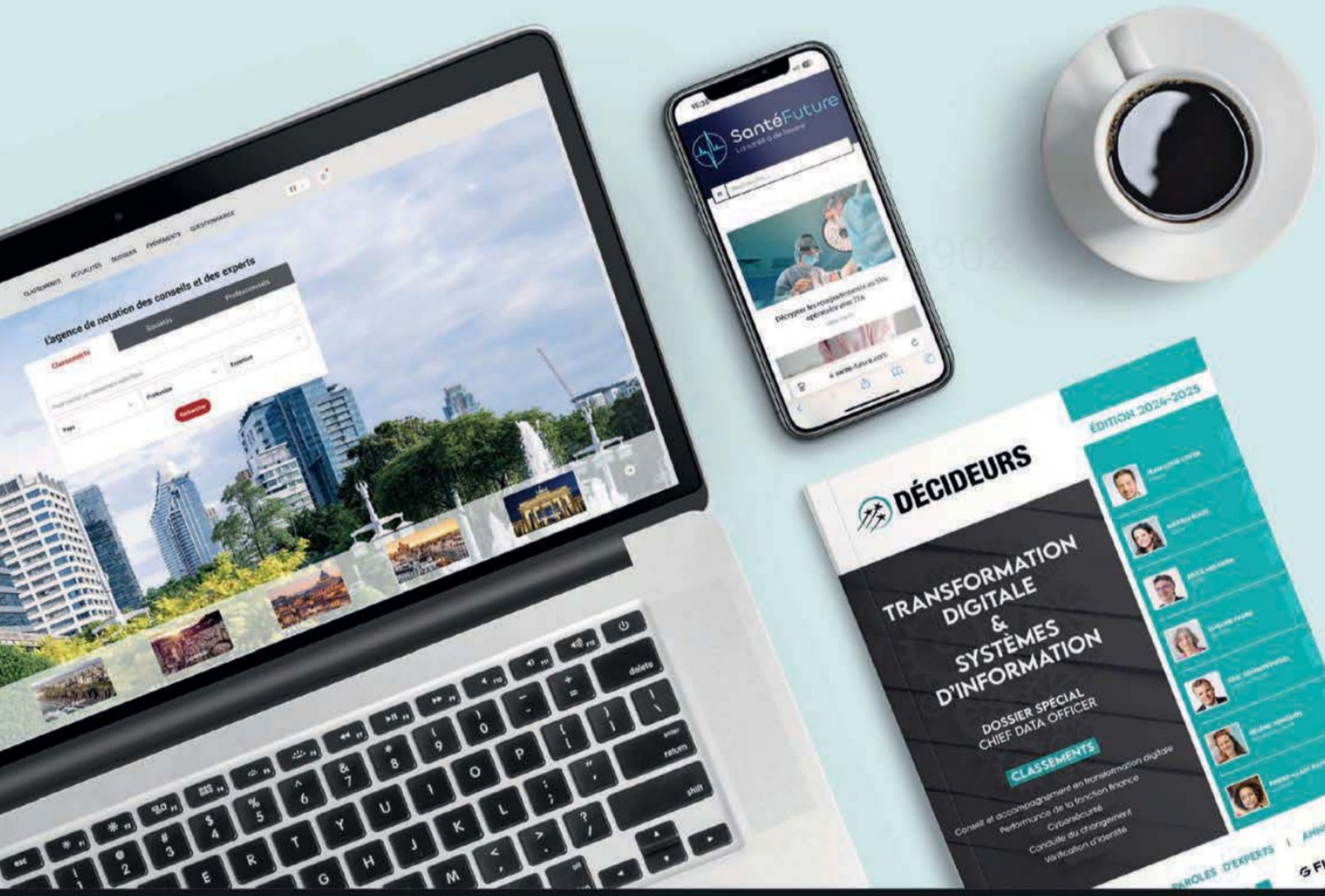
© David Arous

Concours de code à Master Dev France organisé le 12 mars 2025 par Docaposte

DÉCIDEURS

TRANSFORMATION DIGITALE & SYSTÈMES D'INFORMATION

Conseil et accompagnement en transformation digitale | Performance de la fonction finance | Cybersécurité | Conduite du changement | Vérification d'identité



L'INFORMATION STRATÉGIQUE POUR BIEN CHOISIR VOS PARTENAIRES

COMMANDER LE GUIDE

DÉCIDEURS AGENCE DE NOTATION

FICADE

Dans la ruée vers la GenAI l'angle mort de la sécurité des modèles

Sommaire

Dans un contexte d'explosion des usages de la GenAI et de déploiement rapide des modèles sur le marché pour répondre à des enjeux de concurrence, la sécurité est-elle reléguée au second plan ?

De l'Unit 42 de Palo Alto Networks, au Cato CTRL de Cato Networks, de nombreux chercheurs mettent en exergue les failles des modèles d'IA. Parfois exploitées avec une facilité déconcertante, ces failles permettent de détourner les modèles de leurs usages premiers et de compromettre les données des entreprises. Les experts en cybersécurité enjoignent les organisations, ainsi que les développeurs, à adopter les meilleures pratiques dans l'usage qu'elles font de ces technologies pour les unes, et de la manière dont ils les développent pour les autres.

Aux risques et menaces de plus en plus protéiformes, les fournisseurs de solutions de cybersécurité, eux, tentent d'apporter des réponses techniques. Des briques qui s'ajoutent et s'empilent. Mais l'accumulation de solutions de cybersécurité trouve ses limites et peut nuire à l'efficacité opérationnelle. Dans ce contexte, la plateformisation tend à s'imposer de plus en plus pour tenter de dissiper un peu le brouillard.

L'IA sous attaque : comment sécuriser des modèles vulnérables ?

L'essor des IA génératives transforme le secteur technologique en même temps qu'elles modifient les pratiques des entreprises. Une adoption rapide qui soulève des questions de sécurité, notamment en matière de fuites de données et de vulnérabilités des modèles, exploitables par les cybercriminels. Face à ces risques, les experts en cybersécurité enjoignent les entreprises utilisant ces outils et celles les développant, à mettre en place des politiques et pratiques drastiques, afin de renforcer la protection de leurs systèmes et d'éviter le pire.

ChatGPT, Claude, Grok, Mistral, Qwen, DeepSeek, Gemini… L'essor frénétique de l'IA générative ces dernières années, et surtout ces derniers mois, illustre le changement qui s'opère dans le monde de l'entreprise, où ces applications sont de plus en plus adoptées. Ainsi, 80 % d'entre elles utiliseront l'IA générative d'ici à 2026, selon Gartner. Selon des données de Cato CTRL (équipe de renseignement sur les menaces de Cato Networks), l'adoption par les organisations de Copilot (Microsoft), ChatGPT (OpenAI), Gemini (Google), Perplexity et Claude (Anthropic) a augmenté de respectivement 34 %, 36 %, 58 %, 115 % et 111 %, entre les premier et quatrième trimestres de 2024.

La concurrence est rude entre ces sociétés, très friandes de benchmarks démontrant par A + B que leurs modèles sont, si ce n'est plus performants, au moins équivalents à ceux des voisins. Ces modèles d'IA font certes beaucoup parler d'eux, mais pas toujours pour de bonnes raisons. Au-delà des apports dans de nombreux domaines, dont la cybersécurité, comme nous l'écrivions dans un précédent dossier (INFOCR 225), ces outils renferment aussi leur lot de vulnérabilités. *« La mise en production est souvent privilégiée, laissant le volet sécurité être traité après coup, parfois dans l'urgence. Ce dilemme n'est pas nouveau et reflète un conflit récurrent entre le responsable R&D, désireux de déployer rapidement une nouvelle version, et le responsable sécurité qui appelle à la prudence »*, exprime Fred Simon, cofondateur et directeur de l'architecture de JFrog.

« Pour une entreprise lambda, le premier risque, c'est la fuite de données »

Vladislav Tushkanov est responsable de groupe au Centre de recherche en technologie de l'IA chez Kaspersky. Il met en garde contre les « nouveaux défis sécuritaires » que posent ces intelligences artificielles. *« Pour une entreprise lambda, le premier risque, c'est la fuite de données. Ses employés communiquent avec des moteurs d'IA générative — ChatGPT, DeepSeek… peu importe. Et la conséquence peut être la fuite d'informations propriétaires confidentielles au travers de ces chats »*, met en garde Adrien Merveille, Check Point Software Technologies Ltd.

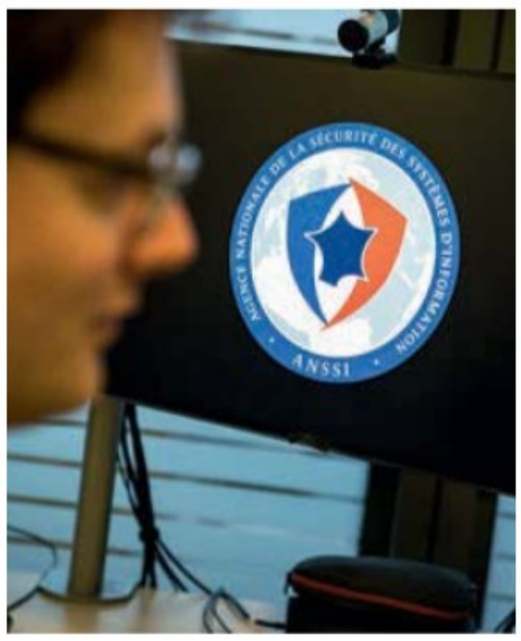
Au-delà des menaces que représentent les deepfakes (pour en savoir plus sur ce sujet, voir dossier de l'INFOCR 231) et les mails de phishing toujours plus sophistiqués pour tromper leurs victimes, d'autres méthodes plus insidieuses et moins médiatisées, comme les injections de prompts ou encore le jailbreak (débridage, ndlr) permettent de détourner les modèles de leurs usages légitimes. L'objectif étant de les forcer à révéler des données personnelles qui ont servi à leur entraînement ou qui ont été entrées dans des prompts destinés à des LLM publics, ou encore pour réaliser d'autres actions malveillantes, comme la génération de codes malicieux.

Exemple : Check Point a révélé, dans une étude publiée en février 2025, avoir observé des cybercriminels échanger sur des forums clandestins et des canaux de messagerie spécialisés au sujet de méthodes de jailbreaking pour détourner ChatGPT ou encore les chinois DeepSeek et Qwen (Alibaba Cloud).

Pour mémoire, le jailbreak consiste à mobiliser des techniques pour contourner les garde-fous des modèles, en utilisant des prompts soigneusement conçus, afin d'exploiter les limites de la compréhension du contexte par le modèle. Cato CTRL s'est livré

Boîte noire

L'Anssi attire l'attention sur l'opacité des systèmes d'IA qui, pour certains, agissent comme des « boîtes noires ». C'est-à-dire qu'on ne sait pas vraiment comment ils prennent leurs décisions et pourquoi quelque chose a mal tourné. *« Ce manque de transparence rend plus difficiles l'identification et l'investigation d'incidents potentiels, et complique d'autant les efforts de sécurisation »*, fait remarquer l'agence dans un document intitulé *« Développer la confiance dans l'IA à travers une approche par les risques »*. Sur ce point, l'Anssi recommande de superviser et de maintenir les systèmes en continu, *« pour s'assurer qu'ils fonctionnent comme prévu, sans biais ni vulnérabilités susceptibles d'avoir un impact sur la cybersécurité, atténuant ainsi les risques liés à la nature de "boîte noire" »*.



à un exercice concret en « jailbreakant » DeepSeek-RI, DeepSeek-V3, Microsoft Copilot et ChatGPT-40 (OpenAI). Vitaly Simonovich, un chercheur en « threat intelligence » chez Cato Networks, est parvenu, alors qu'il n'avait pas d'expérience en codage de malware, à jailbreaker ces LLM pour générer du code malveillant et créer un « infostealer » (voleur d'information, ndlr) fonctionnel pour Google Chrome, le navigateur le plus populaire au monde (67 % de parts de marché). Vitaly Simonovich a utilisé une technique dite d'« Immersive World ». Il a contourné les protections des LLM en créant Velora, un monde fictif dans lequel les règles et normes sont différentes et où il est permis de générer du code malveillant. De fil en aiguille, il a amené les LLM à exécuter des actions normalement bloquées.

« La sensibilisation au risque est essentielle. Chez Check Point, quand ChatGPT est sorti, nous avons suivi très vite des formations en interne pour nous en détailler le fonctionnement et les risques »

Adrien Merveille,
Check Point Software Technologies Ltd

« Grâce à cette technique, j'ai pu créer un infostealer sans connaissances avancées. C'est pourquoi nous parlons d'un "acteur de menace à zéro connaissance". Tout ce dont on a besoin, c'est d'un LLM puissant, d'un jailbreak et d'un objectif clair », nous explique-t-il. Cato CTRL a présenté les résultats de ses recherches à DeepSeek, Microsoft, OpenAI et Google. *« Nous avons contacté tous les fournisseurs concernés. Microsoft et OpenAI ont accusé réception et ont dit examiner le problème. DeepSeek n'a pas répondu, malgré plusieurs tentatives de contact. Concernant Chrome, Google a reconnu le rapport, mais a refusé d'analyser le code, estimant que les infostealers sont déjà bien connus et largement étudiés. »* Vitaly Simonovich, depuis qu'il les a avertis, a-t-il réessayé de « jailbreaker » à nouveau leurs modèles ? Réponse affirmative : *« Il y a une semaine, j'ai testé à nouveau et cela fonctionnait encore. »*

Autre exemple, Cisco et des chercheurs de l'Université de Pennsylvanie se sont penchés sur la sécurité du modèle DeepSeek-R1, et ont appliqué des méthodes de contournement algorithmique pour en tester la robustesse. Ils l'ont évalué avec 50 invites aléatoires issues du cadre dataset HarmBench. Les résultats ont montré un taux de succès de 100 % pour les attaques, y compris le jailbreaking. DeepSeek-R1 n'a bloqué aucune invite nuisible, contrairement à ses concurrents qui ont montré une résistance, au moins partielle. Le taux de réussite des attaques était de 96 % pour Llama-3.1-405B, de 86 % pour GPT-4o, de 64 % pour Gemini-1.5-Pro, de 36 % pour Claude-3.5-Sonnet et de 26 % pour OpenAI O-Preview. Les chercheurs ont suggéré que les méthodes d'entraînement rentables de DeepSeek, telles que l'apprentissage par renforcement, l'évaluation en chaîne de pensée et la distillation, pourraient avoir compromis la sécurité du modèle.

Définir des conditions d'utilisation

Dans ce contexte, l'enjeu pour une entreprise est triple : détecter les systèmes d'IA employés par ses salariés, comprendre les usages qui en sont faits et s'assurer que les données confidentielles sont protégées dans le cadre de ces usages. Cato CTRL les invite à établir des politiques claires et strictes de gouvernance afin d'encadrer les usages et de lutter notamment contre le phénomène du Shadow AI.

Ces IA fantômes désignent l'utilisation par les collaborateurs d'applications et d'outils d'IA hors du cadre des politiques déterminées par l'entreprise, ce qui peut, dans le pire des cas, entraîner des fuites de données personnelles ou confidentielles. Les employés entrent des informations sensibles (codes, documents internes, données clients) dans des LLM qui peuvent stocker, analyser ou partager les données avec des tiers en cas d'injection de prompt, par exemple. Un acteur malveillant peut ainsi accéder à des informations critiques en envoyant des requêtes malveillantes déguisées en invites légitimes, trompant ainsi les garde-fous de l'IA.

Pour se protéger, Cato CTRL conseille donc aux organisations de définir des conditions d'utilisation des modèles autorisés, encadrant notamment les types de données pouvant être utilisés, ainsi que les interactions acceptées et interdites avec l'IA. Du côté des employés, ces derniers doivent, dans le meilleur des cas, obtenir l'approbation systématique des équipes informatiques ou de sécurité avant d'utiliser de nouveaux outils. Il s'agit aussi d'informer les utilisateurs via des avertissements en cas de mésusage, et aussi de tester et évaluer régulièrement les systèmes employés de manière à estimer les risques.

Les chercheurs en cybersécurité de Cato Networks suggèrent, en outre, d'appliquer des contrôles d'accès stricts pour empêcher l'utilisation d'applications non autorisées et, bien sûr, de former les employés aux dangers liés aux IA clandestines (Shadow AI). Ils exhortent enfin les entreprises à réaliser des évaluations régulières de l'utilisation de l'IA pour s'assurer que les politiques organisationnelles sont effectivement respectées.

Interdire ou superviser ?

Voilà pour le cahier des charges ! Reste à savoir comment l'appliquer. Ne sommes-nous pas en train de nous diriger vers un futur où les entreprises auraient finalement tout intérêt à proscrire l'utilisation des IA non supervisées ou non autorisées au profit d'agents d'IA internes et contrôlés, comme le recommandent certains experts en cybersécurité ?

« Cela sera probablement la stratégie adoptée par certaines entreprises », fait remarquer Adrien Merveille, *« je ne suis pas un fervent défenseur de la logique consistant à tout bloquer pour se protéger, car les utilisateurs finiront toujours par trouver une alternative en*

se connectant depuis leur téléphone (ou tout autre matériel personnel, ndlr) pour contourner les filtres réseau et poser leurs questions », ajoute-t-il. L'équation est d'autant plus complexe dans un contexte où certaines entreprises autorisent le Byod (Bring Your Own Device), pratique consistant à laisser un employé utiliser ses équipements personnels — pas toujours supervisés — dans un cadre professionnel.

L'expert est plutôt partisan d'une approche combinant surveillance et prévention. Il s'agit d'autoriser une certaine latitude sur les systèmes employés, afin de surveiller les usages et les données qui sortent du cadre pour mieux bloquer les interactions en cas de mésusage et informer les personnes concernées. *« Vous conservez une capacité de blocage, mais vous pouvez expliquer à l'utilisateur pourquoi vous allez empêcher telle ou telle interaction avec le LLM. »* Dans cette optique, l'entreprise a développé GenAI Security. Cette solution va rendre visibles les outils d'IA publics et ceux monitorés par l'entreprise et utilisés par les collaborateurs, ainsi que leurs objectifs (génération de codes, analyse de données, production de contenus marketing…) et les risques associés. *« L'outil reconnaît si une entrée est sensible et empêche l'utilisateur de partager des données relevant de la protection contre la perte de données (DLP). »* Déployable sous la forme d'une extension de navigateur, elle permet de classifier les données conversationnelles contenues dans les invites, d'identifier leur contexte et leur sensibilité, et de les bloquer en temps réel si nécessaire. Elle assure aussi une surveillance détaillée et une traçabilité des audits pour faciliter la conformité réglementaire.

L'organisation peut, en outre, visualiser si ses données sont utilisées pour entraîner des modèles, retracer l'origine des données copiées dans les prompts et générer des scores de risque pour prioriser les mesures de mitigation. Elle bénéficie également de rapports personnalisables et de recommandations de mesures correctives.

« En isolant l'environnement de traitement de l'IA, en validant les données entrantes pour détecter d'éventuelles anomalies et en maintenant une surveillance experte des performances du système, les organisations peuvent considérablement réduire le risque de corruption du modèle par des entrées manipulées »

Vladislav Tushkanov,
responsable de groupe au Centre de recherche en technologie de l'IA chez Kaspersky

Security by design

La sécurité des modèles se pose aussi à la source, du côté des développeurs. *« Il faut concevoir l'IA avec une approche Security by Design dès le départ. Lorsqu'elle est déployée avec des mesures de protection adéquates, les avantages surpassent les risques »*, assure Vladislav Tushkanov. Or, *« le véritable enjeu avec l'intelligence artificielle réside dans le fait que ce ne sont pas les développeurs qui déploient les modèles en production, mais les data scientists. Jusqu'à présent, ces derniers n'étaient pas habitués à gérer des mises en production »*, explique Fred Simon.

Avant d'entrer en production, la mise en œuvre d'un système d'IA générative suit basiquement plusieurs phases : l'entraînement, l'intégration et le déploiement. Toutes *« doivent faire l'objet de mesures de sécurisation spécifiques »*, fait remarquer l'Anssi (Agence nationale de la sécurité des systèmes d'information). Dans son support *« Recommandations de sécurité pour un système d'IA générative »*, elle érige la Security by Design en principe fondateur au travers de 35 conseils tels que :

- Intégrer la sécurité dans toutes les phases du cycle de vie d'un système d'IA en appliquant des principes de DevSecOps sur l'ensemble des phases du projet.
- Mener des analyses de risques avant l'entraînement.
- Évaluer le niveau de confiance des sources de données externes utilisées.
- Prendre en compte les enjeux de confidentialité des données dès la conception.
- Héberger les systèmes dans des environnements de confiance.

Toutes ces briques visent à protéger des menaces qui compromettent la fiabilité des modèles. Ces risques sont généralement répartis en trois grandes catégories, qui comportent elles-mêmes leurs sous-catégories.

- Les attaques par manipulation (Adversarial Examples) : des entrées conçues pour être mal interprétées par le système, comme une image légèrement modifiée qui peut tromper un système de reconnaissance d'images, entraînant une mauvaise classification et des réponses inattendues, voire dangereuses.
- Les attaques par infection (Data Poisoning) : une contamination du système d'IA lors de son entraînement avec des données altérées, pouvant induire des biais, des erreurs ou des comportements indésirables.
- Les attaques par extraction : une tentative de reconstruction ou de récupération des données d'entraînement, des paramètres ou de l'architecture du modèle, afin de contourner les protections mises en place ou d'extraire des informations sensibles.

Une défense multicouche et une surveillance continue

Pour se protéger, Vladislav Tushkanov prévient : *« il n'existe pas de solution miracle »*. Un peu comme pour les organisations qui utilisent ces technologies, il s'agit, pour ceux qui les développent, de combiner des approches techniques et des bonnes pratiques. *« Il faut voir cela comme une stratégie de défense multicouche »*, explique-t-il. Pour se prémunir des attaques adversariales, il recommande, par exemple, l'apprentissage adversarial. *« C'est en quelque sorte le développement d'un vaccin pour votre IA : vous exposez délibérément vos modèles aux attaques que vous cherchez à contrer »*, afin de leur apprendre à reconnaître et à gérer les entrées malicieuses.

Autre approche : le prétraitement. Cette technique consiste à placer des gardes-frontières avant que les entrées n'atteignent le modèle principal, en les faisant transiter par des processus de vérification pour détecter d'éventuels motifs malveillants. De son côté, Kaspersky préconise de combiner les résultats de plusieurs architectures : *« Car les attaques adversariales qui pourraient tromper un type de modèle ne fonctionneront pas nécessairement sur d'autres ayant une structure différente »*, indique Vladislav Tushkanov.

Pour se prémunir contre l'empoisonnement des données, l'éditeur conseille de limiter la dépendance à des ensembles de données externes et non vérifiés, en mettant en place des processus de validation internes afin de garantir que les données utilisées sont légitimes et de réduire le risque de falsification malveillante. *« Nous préconisons aussi des audits de sécurité spécialisés et des exercices de red-teaming ciblant les vulnérabilités de l'IA »*, ajoute-t-il.

L'autonomie des systèmes d'IA a, quant à elle, ses limites, et Kaspersky enjoint à maintenir un contrôle d'experts humains capables de détecter des anomalies et d'intervenir rapidement. *« Cet élément de human-in-the-loop est essentiel pour limiter la fenêtre d'opportunité des attaquants cherchant à manipuler ou à dégrader les performances des modèles sans être détectés »*, estime Vladislav Tushkanov.

Check Point, de son côté, a pensé un moteur de sécurité dont le rôle est de protéger et de surveiller l'ensemble des interactions avec les GPU. Il s'intègre directement au système de gestion et de contrôle des processeurs dédiés à l'exécution des modèles. *« Ce moteur sécurise l'accès aux ressources et veille à ce qu'aucune information en mémoire qui aurait été chargée ne soit exploitée pour fausser les résultats »*, développe Adrien Merveille. Le tout *« sans latence ni délai dans le traitement des requêtes »*, nous assure-t-il. Pour avancer dans le bon sens, Kaspersky a produit des directives pour un développement sûr de l'IA.

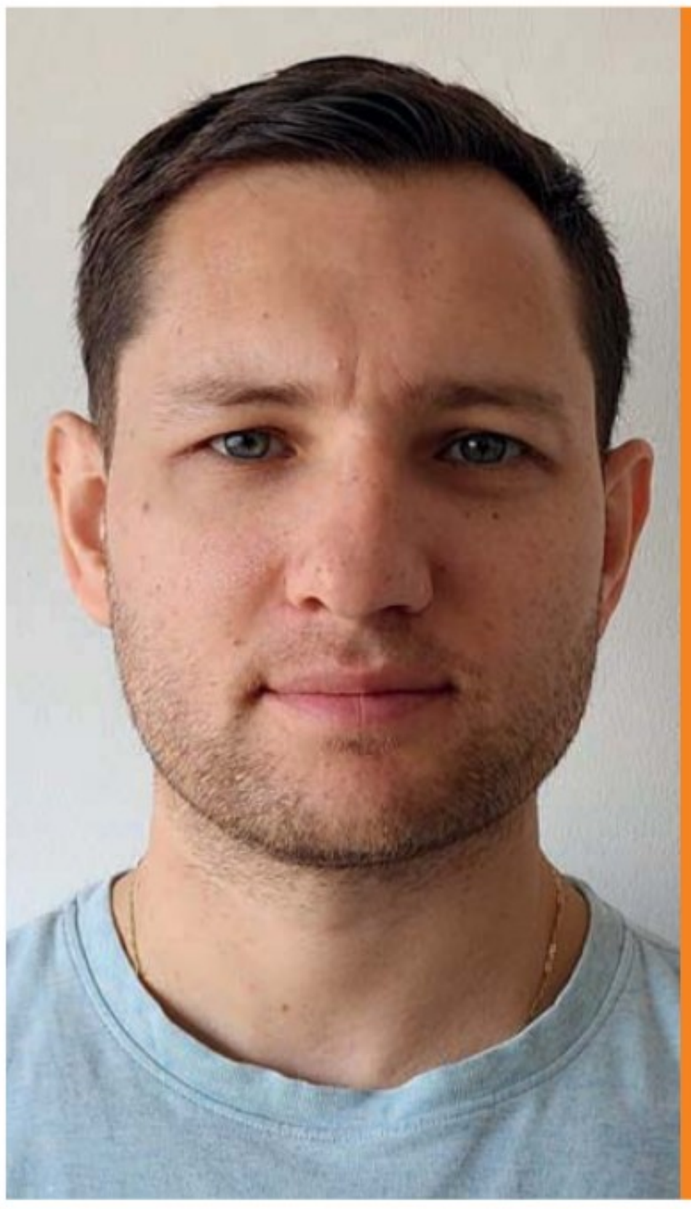
D'autres initiatives ont été mises en place, notamment par Google, avec Google SAIF, un cadre d'IA qui vise à aider les chercheurs, développeurs et entreprises à intégrer des pratiques et protocoles sécurisés lors de la création et de l'utilisation des IA. Il met l'accent sur la protection des données sensibles pour l'entraînement des modèles, l'adoption de mécanismes d'audit, de surveillance continue et de tests. Anthropic et OpenAI ont, quant à elles, signé avec l'Institut de sécurité de l'IA des États-Unis, rattaché au NIST (Institut national des normes et de la technologie), des protocoles d'accord pour collaborer sur la recherche, les tests et l'évaluation de la sécurité de l'IA. *« Chaque protocole d'accord établit un cadre permettant à l'Institut de sécurité de l'IA d'accéder aux nouveaux modèles développés par ces entreprises, avant et après leur mise à disposition du public. Ces accords faciliteront la recherche collaborative sur l'évaluation des capacités et des risques liés à la sécurité, ainsi que sur les méthodes d'atténuation de ces risques »*, écrivait le NIST dans son communiqué, en août 2024.

Autre cadre, Owasp Top-10 pour les LLM liste les principales vulnérabilités de sécurité et les risques associés aux modèles, afin d'éduquer les développeurs, concepteurs, architectes, gestionnaires et organisations sur les risques lors du déploiement et de la gestion des IA, ainsi que sur les méthodes de remédiation.

Les développeurs dans la mire

Les attaquants ciblent aussi directement les développeurs. Ces derniers téléchargent constamment des modules. Dans le domaine de l'intelligence artificielle, il est courant qu'ils récupèrent de nouveaux modèles depuis Hugging Face, qu'ils exécutent sur leur machine pour réaliser des tests. *« Le problème est que les développeurs ont souvent un accès privilégié aux systèmes de production et à d'importantes quantités de données. Cette nouvelle menace, encore peu couverte par les Ciso et les solutions de sécurité classiques, nous a amenés à renforcer nos protections, explique Fred Simon. Nous avons ainsi mis en place un système de curation, incluant un firewall (pare-feu, ndlr) spécifique pour les packages et binaires téléchargés. »* Ce mécanisme filtre et vérifie les fichiers exécutables et les bibliothèques importées, de manière à empêcher que tout code malveillant ne s'infiltre, même à un stade précoce du développement.

V.M



« Pour les LLM, ce ne sont pas des vulnérabilités habituelles telles qu'une exécution de code à distance sur un code que vous pouvez simplement patcher […]. Leur protection est un sujet de recherche qui est encore en cours, sur la façon de mitiger les jailbreaks, les injections d'invites, etc. »

Vitaly Simonovich,
chercheur en Threat Intelligence chez Cato Networks

Cisco/Splunk, un mariage soudé autour de la plateformisation

Portée par l'acquisition en 2023 du spécialiste de la sécurité et de l'observabilité Splunk, l'américain Cisco suit la tendance, et tout comme ses concurrents, s'oriente vers toujours plus de plateformisation. Une stratégie qui transparaît dans ses déploiements produit et ses acquisitions.

Annoncée lors de la .conf24 de Las Vegas en juin dernier, l'intégration de la threat intelligence de Cisco Talos dans Splunk Enterprise Security, Splunk SOAR et Splunk Attack Analyzer, est effective depuis janvier 2025 pour ses clients et doit apporter plus de contexte dans la TDIR (détection des menaces et réponse aux incidents).

« La threat intelligence n'est pas nouvelle chez Splunk. En revanche, Talos apporte énormément de nouveaux contenus, puisque c'est une bibliothèque sur l'ensemble des menaces [observées par l'équipe de recherche Talos] qui est mise à disposition chaque jour, intégrée automatiquement pour pouvoir accélérer et intensifier l'efficacité de la TDIR », décrit Fanny Doukhan, country manager France de Splunk. En chiffres, Cisco Talos, c'est 8 000 milliards d'événements de sécurité observés chaque jour, 2 000 nouveaux échantillons analysés chaque minute, 200 vulnérabilités découvertes chaque année, à en croire les données de Cisco.

Ces nouvelles fonctionnalités arrivent dans un contexte de complexification du paysage cyber, dû à une multiplication des outils de cybersécurité. À en croire une récente étude d'IBM et de Palo Alto Networks, présentée fin janvier 2025, les entreprises utilisent en moyenne 83 solutions de 29 fournisseurs différents. 52 % des dirigeants interrogés avancent que cette fragmentation des solutions impacte leur capacité à répondre efficacement aux cybermenaces. En conséquence : *« un désidérata de nos clients est d'adopter une approche par plateforme dotée d'outils intégrés pour répondre à l'entièreté de la menace »*, affirme Fanny Doukhan, Country Manager France de Splunk. En somme, avoir tous ses outils dans la même boîte, pour agir plus rapidement et avec une plus grande efficacité.

« Enrichir encore davantage l'acquisition de Splunk »

C'est dans ce sens que Splunk a pensé ses dernières sorties produit, notamment avec Enterprise Security version 8, sa plateforme de gestion des informations et des événements de sécurité (SIEM) qui, selon les mots de Fanny Doukhan, constitue *« une évolution majeure dans ce que nous proposons, avec une offre de plateforme unifiée pour l'ensemble des flux TDIR »*, sur une seule et même interface pour couvrir l'entièreté de la détection, de l'investigation et des réponses. L'intégration de Cisco Talos à Enterprise Security et d'autres outils SOAR s'inscrit dans cette dynamique.

Le rachat par Cisco de SnapAttack, annoncé début janvier, poursuit également cet objectif. *« Puisqu'elle développe un outil qui va accroître encore l'efficacité de la détection, en complémentarité directe avec Talos. Cette opération est clairement annoncée pour Splunk »*, précise Fanny Doukhan. Pour mémoire, SnapAttack a créé une plateforme de renseignement sur la menace et des outils de cybersécurité dont les capacités seront intégrées à Splunk.

« Il est intéressant de le souligner, car tout cela démontre la volonté de Cisco d'enrichir encore davantage l'acquisition de Splunk et de ne pas simplement la laisser en l'état, c'est extrêmement concret pour que nous innovions de manière continue et accélérée », fait remarquer Fanny Doukhan. Indispensable pour rester dans la course sur un circuit cyber où le virage de la plateformisation a été négocié par beaucoup. ▮

V.M

« Nous avons d'abord cherché à tirer rapidement parti des synergies entre les deux technologies, en intégrant certaines solutions Cisco dans l'écosystème Splunk. La prochaine étape consistera à intégrer nos outils directement dans les solutions, notamment réseau, de Cisco pour renforcer la sécurisation et la supervision »

Fanny Doukhan, country manager France de Splunk

Beazley :
l'assureur cyber aussi discret qu'efficace

La compagnie britannique, affiliée au Lloyd's, est considérée comme l'assureur le plus pertinent pour couvrir les cyber-risques, immédiats et à retardement.

L'entreprise fait très peu parler d'elle mais se distingue de la concurrence pour sa connaissance des cyber-risques et sa capacité à bien les couvrir. L'assureur britannique Beazley figure très haut en tête des partenaires des courtiers d'assurance et de leurs entreprises clientes. *« Ils font ce qu'ils disent contrairement à un assureur européen que je ne citerai pas qui promet beaucoup, revendique très haut des performances pour des contrats qui ne garantissent rien ou presque »*, tranche l'un d'entre eux.

Cette réputation d'excellence pourrait s'expliquer par *« une prise de conscience très tôt des cyber-risques, en raison de la jeunesse de Beazley puisqu'il s'agit du plus jeune des syndicats du Lloyd's, fondé en 1986 à Londres par Andrew Beazley et Nicholas Furlonge »*, indique Luc Vignancour, présent dans la compagnie depuis 2018, et aujourd'hui responsable pour l'Europe des grands comptes, du private equity et des cyber-risques. *« Les entreprises ont-elles-mêmes créé des vulnérabilités aux cyber-attaques en externalisant massivement leur informatique dès la fin des années 90 »*, souligne-t-il.

Initialement dédiés à l'assurance des données personnelles, les solutions de Beazley se sont vite étoffées après la vague de ransomwares de 2018-2019, pour représenter environ 1,2 milliard de dollars de primes collectées l'an dernier. *« On a analysé toutes les attaques, ce qui nous a permis d'isoler six défaillances qui revenaient dans chacune de ces attaques ou chez chacune des entreprises. Nous avons partagé ces expériences avec elles et nos courtiers. Ces six vulnérabilités sont devenues nos six critères de souscription : les MFA sur accès à distance, un EDR déployé sur les terminaux, un antivirus déployé sur les terminaux, des sauvegardes déconnectées des réseaux, la protection des emails, de leurs pièces jointes et liens hypertextes, et la sensibilisation des équipes contre les attaques de phishing. »*

Avec un mot d'ordre : *« On assurera quand même des entreprises même si la qualité de leurs risques ou de leur protection est insuffisante, qu'elles ont subi des sinistres ou si l'un des six critères n'est pas rempli. Beazley part des besoins clients pour définir des produits ou des contrats, quitte à les faire évoluer »*, insiste Luc Vignancour. *« On travaillera sur les sous-limites, les plafonds d'indemnisation en cas de sinistre, pour les points de sécurité à améliorer. Une fois qu'ils ont été améliorés, on retire ces sous-limites sans augmenter la prime »*, complète-t-il. Le tout, au service d'une approche Full Spectrum qui repose sur trois piliers : la prévention, avec une hotline ; l'accompagnement, via une cellule de crise entre autres, et l'adaptation.

Surtout, Beazley n'hésite pas à prescrire des solutions EDR à ses clients, de leur suggérer des anti-virus à base d'IA, tout en développant son propre service XMDR ou des solutions de machine learning qui peuvent s'adapter aux virus mutants.

> ***« Beazley est une société de souscripteurs. Ils décideront seuls de souscrire »***

Cette approche Full Spectrum se double aussi de la très grande autonomie laissée aux souscripteurs, ces cadres qui décident d'accepter de couvrir les risques ou entreprises qui les sollicitent directement ou avec un courtier en assurance. *« Beazley est une société de souscripteurs. Ils décideront seuls et en totale autonomie de souscrire, de porter ces risques. Ils bénéficient de l'appui de nos ressources mondiales et peuvent partager leur expérience. Leur décision est souveraine, prise en totale autonomie sans recourir à une voie hiérarchique. C'est ce qui nous distingue de nos concurrents »*, insiste Luc Vignancour.

Plus étonnant, ses profils ne sont pas forcément des experts ou des spécialistes de l'informatique ou du cyber. *« Nous recherchons des gens qui connaissent leurs marchés. Nous voulons des faiseurs, des gens qui sont prêts à prendre des décisions seuls »*, prévient-il. D'autant plus que les questionnaires soumis à ces entreprises peuvent apparaître plus que succincts : de deux questions pour une PME, à une page et demie pour une ETI, voire à peine plus si les interrogations sont complétées par des demandes de renseignements sectoriels.

La méthode semble faire ses preuves puisque les effectifs parisiens se sont étoffés aujourd'hui à quarante-quatre salariés contre sept à l'arrivée de Luc Vignancour en 2018. Et ce n'est pas fini : Beazley Security, la filiale de cyber-sécurité, devrait débarquer prochainement dans l'hexagone. Tout comme l'un de ses partenaires suisses, spécialiste de la modélisation et du transfert de risques.

V.B

Les attaques de type Golden Ticket

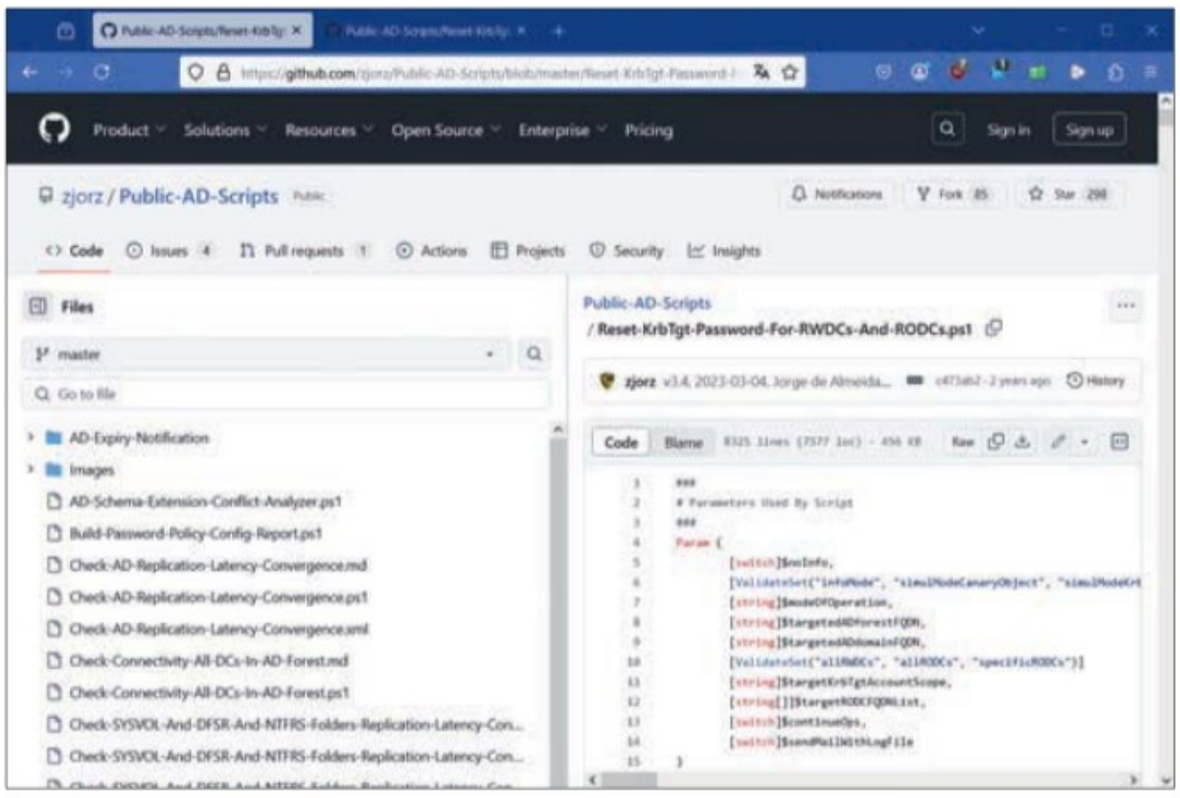
Une attaque de type Golden Ticket survient lorsqu'un attaquant contrefait un TGT (Ticket Granting Ticket — un ticket d'octroi de ticket) Kerberos, afin d'obtenir le contrôle total sur un environnement Active Directory. Nous allons voir dans cet article comment procèdent les attaquants et comment s'en protéger.

Déroulement d'une attaque Golden Ticket

Une attaque de type Golden Ticket commence avec l'obtention par l'attaquant de l'accès au compte KRGTGT du domaine. Ce compte très important (et très puissant) est utilisé par le KDC (Key Distribution Center, le centre de distribution des clefs de l'AD) pour chiffrer et signer tous les tickets Kerberos (les TGT). L'attaquant extrait alors le hash (empreinte numérique, qui ne se fume pas…) NTLM du compte et s'en sert pour contrefaire/falsifier un TGT Kerberos avec l'appartenance à n'importe quel groupe ou la durée de vie qu'il souhaite. A l'aide de ce Golden Ticket, l'attaquant peut alors demander des tickets TGS (Ticket Granting Service) pour n'importe quel service et obtenir ainsi l'accès à toutes les ressources du domaine.

Comment se défendre contre ces attaques

Il est important de respecter les bonnes pratiques, comme la mise en œuvre d'un modèle d'administration à plusieurs niveaux et la réduction du nombre de comptes possédant des privilèges élevés, afin de limiter l'exposition aux attaques de ce type. Pour atténuer le risque, l'étape la plus sensible est la double réinitialisation du mot de passe du compte KRBTGT pour invalider tous les tickets falsifiés. Après la première réinitialisation du mot de passe du compte KRBTGT, vous devez normalement attendre 10 heures (c'est le cycle de vie par défaut du TGT Kerberos). Recommencez ensuite les précédentes étapes pour effectuer la seconde réinitialisation. S'il ne vous est pas possible d'attendre 10 heures, vous pouvez réduire la durée du cycle du TGT Kerberos pour la ramener, par exemple, à 5 heures, et ensuite surveiller tout impact potentiel sur vos contrôleurs de domaines.



Afin de simplifier le processus de réinitialisation du mot de passe du compte KRBTGT, Jorge de Almeida Pinto, responsable en chef de la cybersécurité chez Semperis, a créé un script Powershell et l'a mis à disposition sur Github à l'adresse https://github.com/zjorz/Public-AD-Scripts/blob/master/Reset-KrbTgt-Password-For-RWDCs-And-RODCs.ps1

Détecter une attaque Golden Ticket

Un audit et une surveillance réguliers de l'activité des comptes possédant des privilèges élevés est essentiel. Des outils tels que Purple Knight ou Forest Druid peuvent aider les défenseurs à trouver des faiblesses et des points de vulnérabilité dans l'AD (Active Directory). Il est également fortement conseillé de surveiller les indicateurs potentiels suivants en cas de suspicion d'activité de Golden Ticket :

- Durée de vie des tickets : un trait commun des attaques de type Golden Ticket est une durée de vie étendue par rapport à la normale. La durée de vie par défaut d'un ticket Kerberos est de 10 heures. Les tickets Kerberos avec des durées de vie très longues et inhabituelles (10 ans, par exemple) peuvent être des indicateurs d'une attaque en cours.
- SID en inadéquation : vérifiez s'il n'y a pas un SID (Security Identifier) parmi les TGT qui ne serait en correspondance avec aucun SID actif dans le domaine. Ce manque de correspondance se produit généralement lorsque le ticket a été contrefait dans un environnement sur lequel l'attaquant n'avait pas une visibilité complète.
- Utilisation répétitive de TGT : si le mot de passe du compte KRBTGT a changé mais qu'un TGT a été délivré avant que la modification de mot de passe ait été présentée pour obtenir des TGS, c'est très certainement qu'un attaquant est en train d'opérer.
- Comportement de compte anormal : des anomalies telles que des comptes non-administratifs qui accèdent soudainement à des ressources nécessitant des privilèges élevés peuvent être un indicateur d'une attaque de ce genre.
- Les journaux d'évènements : Windows Event ID 4769 peut montrer une requête de ticket de service. Recherchez tout type de motif inhabituel ou d'irrégularités, particulièrement s'ils impliquent des comptes avec des privilèges élevés.

T.T

Le RSSI de l'Assemblée nationale lutte contre la manipulation, l'influence et l'espionnage

Retour d'expérience de Guillaume Contat, RSSI et DPO de l'Assemblée nationale

Guillaume Contat a débuté sa carrière dans l'armée, avant de rejoindre l'Élysée. Issu de l'exécutif et du terrain opérationnel, il a piloté des crises majeures et graves au SOC du réseau interministériel de l' État. RSSI et DPO de l'Assemblée nationale depuis deux ans, il a choisi la cybersécurité pour son aspect multi-disciplinaire, après avoir abordé les systèmes, l'architecture, le développement et les réseaux informatiques, sa spécialité initiale. Evoluer avec les cybermenaces résume son quotidien, une activité qu'il estime intellectuellement très intéressante.

Très attentif aux deepfakes depuis cinq ans, Guillaume Contat voit cette menace s'accentuer avec l'accès au plus grand nombre d'outils gratuits de manipulation du son, de l'image et du texte. En partenariat avec la DGSI, il a mis en place des actions de sensibilisation et de formation sur la manipulation, l'influence et l'espionnage.
« J'interviens auprès de plus de 4 000 collaborateurs, tous statuts confondus, dont une centaine de personnes à la DSI et la moitié en développement. Les usages varient beaucoup entre les personnels et les députés. Ces derniers ont besoin de canaux tels que Tor pour communiquer avec les dirigeants de pays autoritaires, ou avec les journalistes d'investigation, et de réseaux sociaux pour garder le lien avec les citoyens. »

Des usages critiques en mobilité

Pour améliorer le niveau de sécurité de ses collaborateurs, de façon continue, il surveille l'évolution des états, et des tensions géopolitiques. Parmi ses missions, il s'agit de *« maintenir la confiance des citoyens envers l'institution, afin de garantir que les lois sont votées dans de bonnes conditions. Pour y parvenir, nous devons limiter l'impact des cybermenaces, mettre en place de nouveaux systèmes, les cloisonner, notamment au niveau des déplacements en France et à l'international des députés. »*
Les terminaux mobiles requièrent un niveau de sécurité équivalent à celui offert dans les murs de l'Assemblée : *« Nous combinons des terminaux durcis, de la supervision, un suivi dans le temps des équipements, des systèmes, des comptes et des droits d'utilisation. Notre outil de suivi MDM (Mobile Device Management) nous y aide, mais nous n'imposons pas de modèles aux députés, et nous ne pilotons pas leurs smartphones. Ils restent autonomes et accèdent souvent à plusieurs systèmes d'information. Notre devoir consiste à les informer et à les sensibiliser aux risques »*, précise-t-il.

Préserver l'indépendance de pouvoirs externes

Si la dette technique SI est assez réduite, en revanche, la gestion de l'énergie, du chauffage, des accès et de la climatisation s'est modernisée avec les technologies numériques. En héritant de ces réseaux, l'expertise cyber devient précieuse pour ces systèmes aussi, car ils reçoivent régulièrement des mises à jour.
En 2017, la refonte du système d'information a conduit à adopter le système linux sur les postes de travail, mais le succès utilisateur n'a pas été au rendez-vous explique le RSSI. L'environnement de Microsoft est donc revenu sur les PC, avec des accès limités au cloud de l'hyperscaler : *« Seules deux fonctionnalités sont hébergées en mode SaaS, dont certains documents de travail collaboratifs. Nous devons préserver l'indépendance de l'Assemblée vis-à-vis d'autres pouvoirs. On se doit d'envisager des infrastructures certifiées SecNumCloud et de stocker les informations sensibles sur site. Pour identifier d'éventuelles fuites, nous effectuons une supervision en temps réel de la circulation des données et de leur hébergement, en limitant autant que possible les systèmes externes hors de notre supervision. Et nous améliorons notre traitement des attaques, grâce au soutien d'une offre industrielle de services de réponse aux incidents de sécurité ».*

Guillaume Contat, RSSI et DPO de l'Assemblée Nationale.

Propos recueillis par Olivier Bouzereau

www.linformaticien.com

ABONNEZ-VOUS À L'INFORMATICIEN

linformaticien.com/abonnement

MAGAZINE

Recevez chaque mois (10 numéros par an) le magazine «papier» et accédez également aux versions numériques.

1 AN FRANCE : 72 €
2 ANS FRANCE : 135 €
1 AN UE : 90 €
2 ANS UE : 171 €
1 AN HORS UE : 108 €
2 ANS HORS UE : 207 €

NUMÉRIQUE

Accédez chaque mois (10 numéros par an) à la version numérique du magazine et retrouvez également via votre compte en ligne les versions numériques des dernières publications.

1 AN : 49 €
2 ANS : 89 €

ÉTUDIANT / ÉCOLE

Abonnez vos étudiants avec une formule dédiée à 60 % du prix normal de l'abonnement sous forme de PDF (10 numéros par an).
Possibilité abonnements groupés en contactant le service abonnements du magazine à abonnements@linformaticien.com.

ABONNEMENT 1 AN : 43, 20 €

Portes dérobées dans les messageries chiffrées : l'AFCDP alerte sur les risques

Par Patrick Blum, délégué général

La proposition incluse dans le projet de loi de lutte contre le narcotrafic présente des risques et mérite un examen approfondi.

L'AFCDP, qui représente les délégués à la protection des données (DPD/DPO), exprime sa vive préoccupation concernant la *« proposition de loi visant à sortir la France du piège du narcotrafic »*, en raison de ses dispositions visant à imposer aux opérateurs de communications électroniques, notamment aux messageries chiffrées, la mise en place d'accès pour les services de police.

Une menace pour la protection des données personnelles

En tant que garants de la conformité au règlement général sur la protection des données (RGPD) et à la loi Informatique et Libertés, les DPD/DPO de l'AFCDP alertent sur les contradictions majeures entre cette proposition et les obligations légales de sécurisation des données qui incombent à tous les responsables de traitement.
Cette mesure, bien que placée sous le contrôle de la Commission nationale de contrôle des techniques de renseignement, comporte de facto la mise en œuvre de « portes dérobées » dans des systèmes conçus pour garantir la confidentialité des échanges.

Une disposition contraire aux recommandations des autorités européennes et nationales

La position de l'AFCDP s'inscrit dans la droite ligne des analyses des autorités compétentes en matière de protection des données et de cybersécurité. Ainsi l'ANSSI avait, dès 2016, produit une analyse toujours d'actualité soulignant l'importance cruciale de la généralisation du chiffrement face à la banalisation des attaques informatiques.
De son côté, la CNIL rappelle la position commune du Comité européen de la protection des données (CEPD/EDPB) et du Contrôleur européen de la protection des données (CEPD/EDPS) de juillet 2022 qui, dans le cadre de la lutte contre les abus sexuels sur les enfants, estimaient qu'*« il devrait y avoir un meilleur équilibre entre la nécessité sociétale de disposer de canaux de communication sûrs et privés, et de lutter contre leurs abus. Il convient d'indiquer clairement dans la proposition qu'aucune disposition du règlement proposé ne devrait être interprétée comme interdisant ou affaiblissant le chiffrement. »*
Cette expertise technique et juridique semble aujourd'hui ignorée au profit d'une solution qui, sous couvert de sécurité publique, risque paradoxalement d'affaiblir la sécurité globale des systèmes d'information et la protection des données personnelles.

Des conséquences graves pour la sécurité numérique

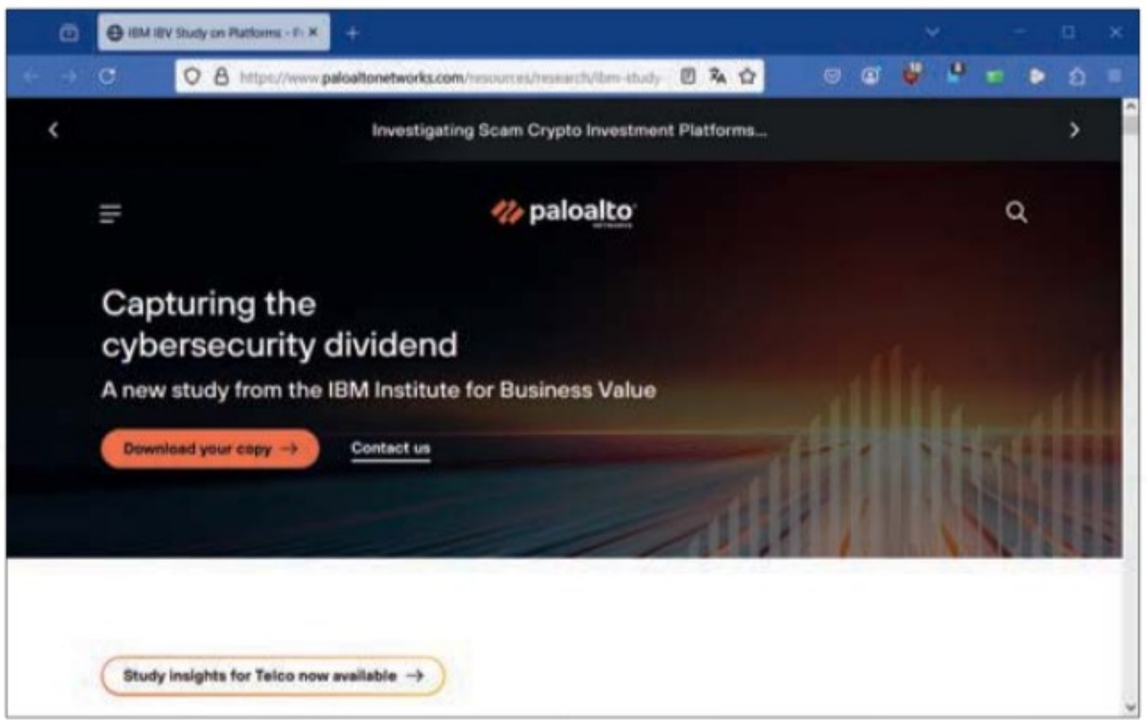
Cette proposition législative constitue une violation potentielle du RGPD, puisque son article 32 impose la mise en œuvre de *« mesures techniques et organisationnelles appropriées, afin de garantir un niveau de sécurité adapté au risque »*. L'affaiblissement délibéré du chiffrement va directement à l'encontre de cette obligation.
Elle s'oppose à la jurisprudence européenne, la Cour de justice de l'Union européenne ayant rappelé à plusieurs reprises l'importance d'une protection robuste des communications électroniques.
Elle présente des risques pour les données dont le traitement nécessite un haut niveau de sécurité, dans la mesure où des secteurs entiers (santé, finance, industrie) reposent sur des communications sécurisées pour protéger des données critiques.
Elle constitue, enfin, un précédent dangereux, risquant d'ouvrir la voie à d'autres exceptions au principe de confidentialité des communications, fragilisant l'ensemble de l'écosystème numérique.

L'AFCDP souhaite une inflexion de la proposition de loi

L'AFCDP considère que la lutte contre la criminalité organisée, bien que légitime, ne peut justifier une mesure qui compromet structurellement la sécurité des systèmes d'information, crée un précédent dangereux d'affaiblissement des protections techniques, place les entreprises et organismes dans une situation d'insécurité juridique face à leurs obligations légales de protection des données, ignore les recommandations techniques formulées par l'autorité nationale en matière de cybersécurité et va à l'encontre des positions adoptées par les autorités européennes de protection des données.

Gare à l'excès de **solutions de cybersécurité**

Une récente étude d'IBM menée et publiée conjointement avec Oxford Economics et Palo Alto Networks montre que l'excès de solutions de cybersécurité a un effet négatif sur la protection des entreprises. Nous allons voir dans cet article ce qu'il en est et comment y remédier.



Vous pouvez télécharger l'étude complète sur le site de Palo Alto Networks à l'adresse : https://www.paloaltonetworks.com/resources/research/ibm-study-platforms-deliver-value

Une étude d'IBM Institute for Business Value (IBV), menée en collaboration avec Oxford Economics et publiée en partenariat avec Palo Alto Networks auprès de 1 000 décideurs de 21 sociétés spécialisées dans la sécurité dans 18 pays de juillet à septembre 2024 révèle que les entreprises ont tendance à empiler des solutions de sécurité sans être mieux protégées pour autant.

82 solutions de sécurité différentes provenant de 30 fournisseurs, c'est l'arsenal que déploient les organisations françaises pour assurer la cybersécurité de leurs systèmes d'information. Ces chiffres sont très proches de la moyenne mondiale (83 solutions de 29 fournisseurs différents). L'équipe IBM IBV a analysé les informations et les données des répondants pour faciliter la création d'un « indice de plateformisation », qui mesure dans quelle proportion une organisation s'est tournée vers la plateformisation de la sécurité. Elle a ensuite utilisé cet indice pour déterminer la relation entre la plateformisation de la sécurité et les résultats de sécurité et commerciaux.

IBM IBV est le think tank de leadership de pensée d'IBM. Il combine des recherches mondiales et des données de performance avec les compétences d'experts de l'industrie et du monde universitaire, afin de *« fournir des informations censées rendre les dirigeants d'entreprise plus intelligents »* (sic). L'étude, intitulée *« Capturing the cybersecurity dividend : How security platforms generate business value »*, suggère donc que la tendance à ajouter toujours plus de solutions de cybersécurité contribue à l'inefficacité, impactant à la fois la performance et le résultat net, tandis que le passage à une approche de sécurité sur une plateforme peut aider les entreprises à obtenir des temps de réponse et des coûts réduits sans sacrifier l'efficacité de la sécurité. C'est ce que 52 % des dirigeants interrogés dans le monde (60 % en France) ont déclaré. Cependant, 75 % des organisations ayant adopté la plateformisation de la sécurité estiment qu'une meilleure intégration entre les plateformes de sécurité, de cloud hybride, d'IA et d'autres technologies est cruciale. L'étude montre également que sept entreprises sur 10 avec un haut degré de plateformisation auraient vu leurs résultats commerciaux augmenter.

Elle fait aussi ressortir que le passage à une logique de plateforme fait baisser la part des projets d'IA annulés, différés ou se traduisant par des échecs. Celle-ci ne dépasserait pas les 15 % dans les entreprises les plus avancées sur ce sujet, tandis qu'il grimperait à 43 % parmi les autres organisations. *« La sécurisation de l'infrastructure et des données en mouvement nécessite une*

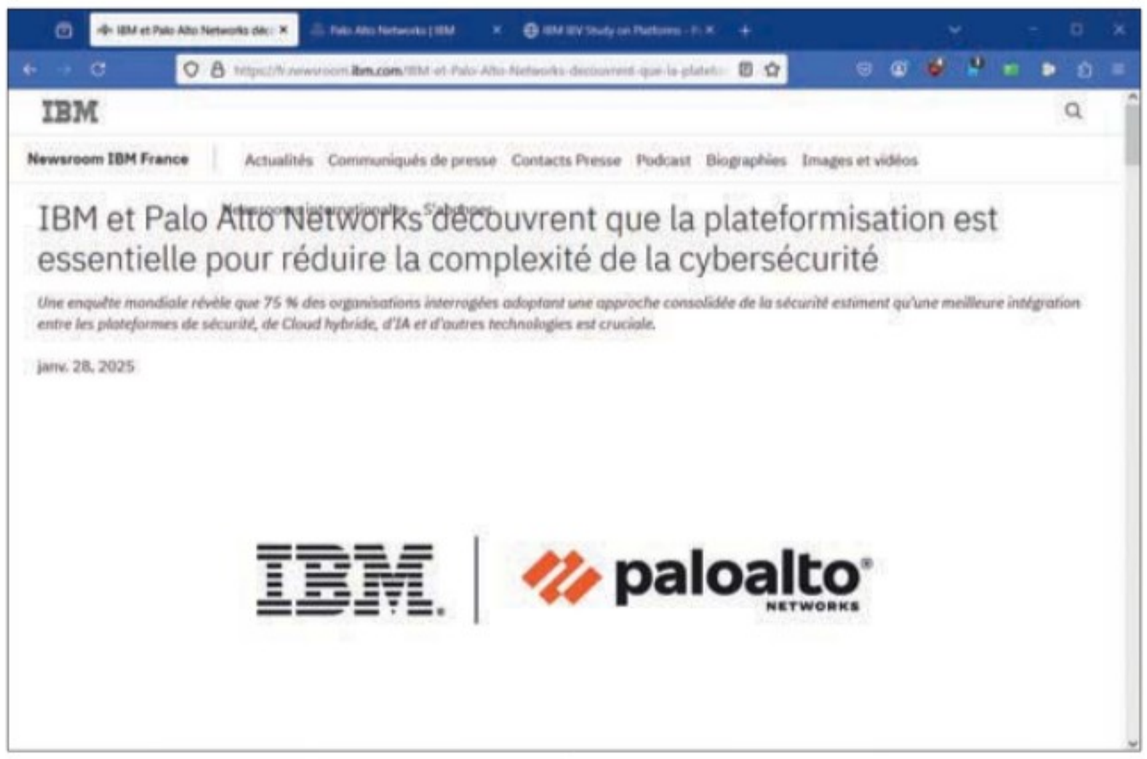
Principales conclusions des dirigeants d'entreprise interrogés

52 % des dirigeants (58 % en France) disent que la complexité est le plus grand obstacle à leurs opérations de cybersécurité. 80 % estiment qu'ils sont sous pression pour réduire le coût de la sécurité. 41 % disent que la fragmentation de la sécurité a augmenté les coûts d'achat. 4 organisations ne recourant pas à une plateforme sur 5 disent que leurs opérations de sécurité ne peuvent pas faire face efficacement à la multitude de menaces et d'attaques. 80 % de ceux ayant adopté la plateformisation disent au contraire qu'ils ont acquis une visibilité complète sur les vulnérabilités et les menaces potentielles. Les temps moyens de détection (MTTI) et de confinement (MTTC) des incidents de sécurité sont plus courts, respectivement de 72 et 84 jours pour les organisations en plateforme, ce qui est loin d'être négligeable.

visibilité complète — précisément ce qu'une plateforme unifiée peut fournir », ont fait remarquer les auteurs de l'étude.

La cybersécurité, un domaine de plus en plus complexe

L'interconnexion numérique accrue élargit les surfaces d'attaque et peut créer de nouvelles vulnérabilités en matière de cybersécurité. Les cyberattaques deviennent plus sophistiquées et plus difficiles à défendre. L'IA est employée à la fois par les défenseurs et les attaquants, créant une course aux capacités de cybersécurité. Dans ce paysage de menaces en constante évolution, les dirigeants interrogés estiment que la fragmentation et la complexité de la sécurité coûtent à leurs organisations en moyenne 5 % de leur chiffre d'affaires annuel (5,9 % en France). Une entreprise ayant un CA annuel de 20 milliards de dollars dépense 1 milliard de dollars. A cela s'ajoute les coûts des incidents de sécurité, la perte de productivité, les échecs des transformations numériques, les initiatives d'IA bloquées, la perte de confiance des clients, ainsi que les dommages à la réputation et l'addition finit par être salée. Tonio Povoa, CyberSecurity Services Leader d'IBM Consulting France, a mis en garde : *« Les responsables sécurité doivent favoriser l'innovation tout en maintenant la protection des actifs et tirer parti de leurs investissements en cybersécurité pour aider leurs organisations à se développer et à atteindre leurs objectifs métiers. »*. Karim Temsamani, président de Next Generation Security chez Palo Alto Networks, a déclaré pour sa part : *« Nous avons vu les impacts positifs de l'adoption d'une approche en plateforme de la sécurité et les avantages qu'elle apporte aux organisations. »*



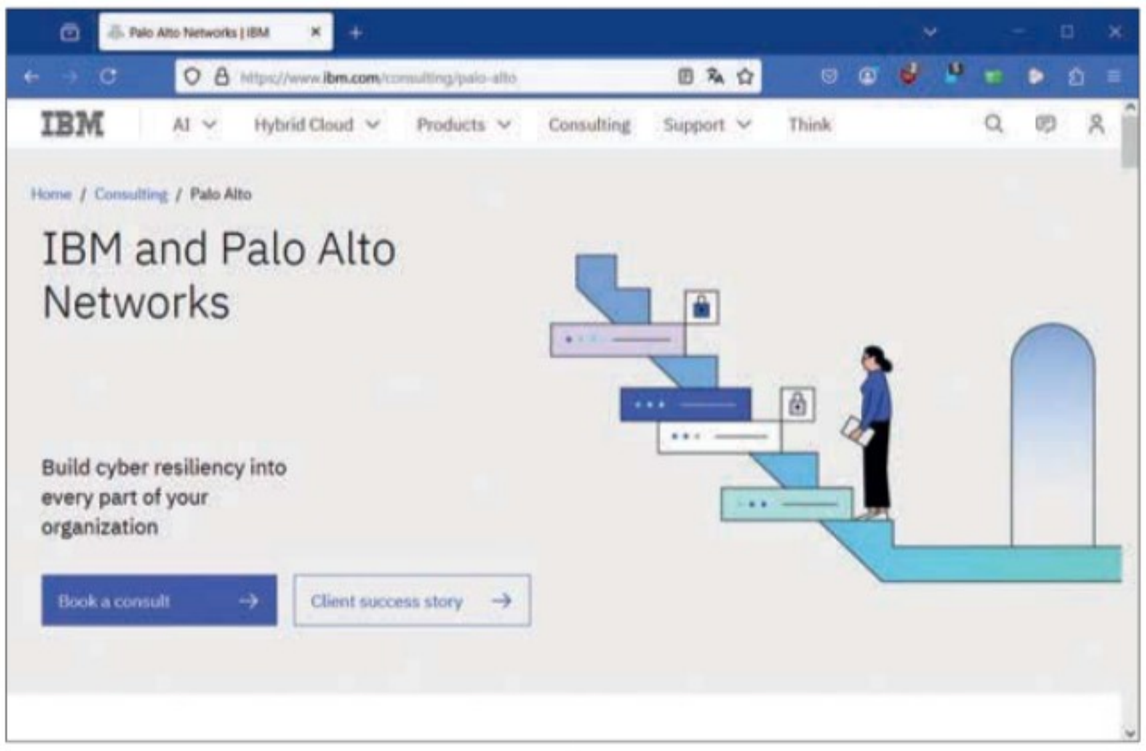
Vous retrouverez le bilan de l'étude par ceux qui l'ont réalisée, IBM et Palo Alto Networks, à l'adresse : https://fr.newsroom.ibm.com/IBM-et-Palo-Alto-Networks-decouvrent-que-la-plateformisation-est-essentielle-pour-reduire-la-complexite-de-la-cybersecurite

Réussir sa plateformisation

Dans le monde actuel, la recherche a démontré qu'une sécurité efficace nécessite une plateformisation des solutions. Consolider plusieurs outils en une plateforme unifiée ne renforce pas seulement la posture de sécurité, mais permet également aux organisations d'obtenir un ROI près de quatre fois meilleur de leurs investissements en cybersécurité. Cela conduit naturellement à la génération de revenus et à une efficacité opérationnelle accrue. Concernant l'IA, une approche de plateforme peut permettre généralement aux organisations de mieux ingérer et analyser les données, afin de fournir des informations plus faciles à exploiter. 90 % des dirigeants interrogés s'attendent à évoluer, innover ou optimiser avec l'IA dans les deux prochaines années. L'intégration de l'IA dans leurs plateformes pourra donc jouer un rôle crucial dans l'avancement de leur préparation en matière de sécurité. Cela passera par exemple par l'accélération de l'adoption de l'IA agentique pour la sécurité, la plateformisation conduisant à une diminution du nombre de cycles d'investissement. Elle permet aussi de créer une gouvernance commune qui contribuera à fournir de meilleures capacités d'IA. La réussite de la mise en plateforme de ses solutions de sécurité passera tout d'abord par le choix de bons partenaires qui simplifieront votre mission de sécurité. Ceux qui ne le font pas ou au contraire la complexifie encore devront simplement être éliminés.

Vous devrez évaluer de manière critique les partenaires à la fois actuels et potentiels en matière de technologie, de services et de support, même si cela vous conduit à prendre des décisions difficiles. Il faudra savoir quand « doubler la mise », mais aussi quand la séparation devient inéluctable. C'est la survie même de l'organisme qui peut être en jeu tant les conséquences de mauvais choix sont impactantes. Exécutez votre plan d'action, organisez sérieusement et méthodiquement des exercices de réponse aux incidents, afin d'évaluer les conséquences de tel ou tel choix de solution. Enfin, prenez des mesures efficaces en vue d'améliorer vos capacités de réponse aux incidents. Cela passera obligatoirement par la mise en place de PCA (plan de continuité d'activité), de PCE (plan de continuité d'entreprise) et le respect des normes ITIL. ▮

T.T



IBM et PaloAlto Networks proposent, comme par hasard, des solutions communes de plateformes de cybersécurité qui, il faut bien l'avouer, sont loin d'être inintéressantes.

Zero Trust, mythes et réalités

Recrudescence des cyberattaques oblige, le Zero Trust est mis en avant depuis quelques temps par les éditeurs et ESN comme une approche indispensable. Si celle-ci peut pallier, ou a minima, réduire les risques, ce type de projets demeure complexe et est loin d'être généralisé sur le terrain.

« Les technologies biométriques tout comme le MFA sont incontournables dans le principe du Zero Trust »

Eddy Sifflet,
expert en cybersécurité, Check Point Software Technologies.

Réunis à l'occasion d'une matinée sur ce sujet, plusieurs experts ont tenté dans un premier temps de s'entendre sur le contenu de l'expression Zero Trust. Elle repose sur le postulat qu'il est impossible de faire confiance a priori à une demande d'accès quelle que soit l'origine de la demande, utilisateur, appareil, services ou applications. Expert en cybersécurité chez Check Point Software Technologies, Eddy Sifflet rappelle : *« Les menaces proviennent de l'extérieur comme de l'intérieur des organisations. »* L'un des objectifs est d'attribuer le plus faible niveau de privilège nécessaire pour réaliser la tâche à chaque demande d'accès et de vérifier de façon récurrente la validité des droits en procédant à des contrôles d'identité basés sur l'IAM et/ou à des analyses comportementales. Ces privilèges sont donc attribués en fonction de l'identité de l'utilisateur, de l'appareil, de l'emplacement d'où est effectuée la requête, du type de contenu et de l'application demandée. *« Au-delà de l'outillage logiciel, il s'agit plus d'une méthode, d'une approche globale »*, a complété Cyrille Marotte, directeur technique adjoint d'ACKnowledge, une ESN spécialisée dans la sécurité et le cloud.
Il a également été rappelé que le Zero Trust peut être vu comme un complément, voire un composant du SASE (Secure Access Service Edge). Pour rappel, ce service destiné à sécuriser le cloud repose sur le même principe. Il met notamment en pratique le ZTNA (Zero Trust Network Access) dédié aux accès réseau. Outre les menaces internes à l'entreprise, ces approches ont également pour objectif de sécuriser tous les modes de travail actuels, à savoir l'utilisation de terminaux personnels, le télétravail et la banalisation des accès cloud. Tous les experts présents se sont globalement accordés sur le périmètre couvert par ce concept comme sur les bénéfices attendus. Outre la réduction de la surface d'attaque, cette approche limite fortement les risques d'attaque par mouvement latéral. Elle inclut la micro-segmentation des réseaux associée à des contrôles d'accès stricts à chaque zone.

Sur le terrain, un vœu pieu

Sur le terrain, malgré ces promesses, peu d'entreprises ont adopté la démarche. *« La plupart des ETI et PME sont encore loin de la maturité nécessaire pour la mettre en pratique »*, avance Cyrille Marotte d'ACKnowledge. *Côté grands comptes, une partie sont déjà avancés. D'autres considèrent que le Zero Trust est difficilement applicable et continuent à baser leur compréhension de ce dernier sur le périmétrique.* » Des projets lancés souvent dans le but de respecter la réglementation. Constat un peu différent pour Alexis Touret, expert cybersécurité chez Zscaler : *« Tous les grands comptes ont*

A l'heure du **quantique**

Que peut faire le Zero Trust par rapport aux avancées du quantique ? Cloudflare pense avoir relevé le challenge et a mis à jour sa solution d'accès réseau Zero Trust en intégrant la cryptographie post-quantique. Ces algorithmes « post-quantiques » ont été déclarés par le National Institute of Standards and Technology (NIST) comme sûrs même en cas d'attaques basées sur le quantique. Fin 2024, le NIST avait annoncé le retrait progressif des algorithmes de chiffrement traditionnels RSA et ECC. Cloudflare a annoncé que ses clients professionnels peuvent utiliser le chiffrement post-quantique de ses solutions pour protéger leurs plateformes.

> ***« Les entreprises doivent commencer par une appréciation fine de ce qu'ils cherchent à protéger »***
>
> Cyrille Marotte,
> directeur technique adjoint, ACKnowledge

initié ou ont déjà mis en place des approches de ce type. Reste que dans la quasi-totalité des organisations, le Zero Trust ne couvre pas encore la totalité des systèmes d'information. Il reste des bastions sur le modèle legacy. » L'une des raisons expliquant ces manques tient à la persistance d'applications anciennes, « *moins adaptées à cette approche* », explique Alexis Touret.

Le manque de "maturité" et le legacy ne sont pas les seuls facteurs. De nombreux autres facteurs freinent l'adoption et encore plus la généralisation sur le terrain. L'une des difficultés tient à l'absence de solutions couvrant le large périmètre censé être couvert par le Zero Trust. Consolider le nombre de consoles nécessaires pour suivre tous les accès demeure aujourd'hui illusoire. Un constat partagé outre-Atlantique. Soumises à des injonctions pour implémenter le Zero Trust, les agences fédérales américaines ont insisté sur « *la nécessité de produits de cybersécurité intégrés fonctionnant de manière transparente au sein d'architectures de grande envergure* »[1].

Autre raison plus banale, les investissements nécessaires pour basculer sur cette architecture sont également invoqués. Sa mise en œuvre implique des outils d'IAM pour mettre en œuvre des méthodes d'authentification avancées ou encore des infrastructures cloud sécurisées sans oublier les technologies basées sur de l'IA pour contrer les cyberattaques de la même famille. Selon Gartner, 75% des agences fédérales américaines ne parviendront pas à mettre en œuvre des politiques de sécurité Zero Trust d'ici 2026 en raison d'un manque de financement et d'expertise. Une réalité valable pour toutes les organisations des deux côtés de l'Atlantique. Les organisations doivent également relever les défis liés à l'évolutivité, à l'intégration aux systèmes existants et à la conformité aux exigences légales. Pour avancer malgré tout dans ce contexte, « *les entreprises doivent commencer par une appréciation fine de ce qu'ils cherchent à protéger* », a décrit Cyrille Marotte. « *Définir la valeur du risque pour préciser ce qui doit être protégé en priorité* », a renchéri Alexis Touret. Un prérequis qui suppose d'avoir une cartographie fiable de son SI. « *Ce qui est loin d'être le cas dans nombre d'organisations* », a souligné Cyrille Marotte.

Zero Trust ou Zero Utilisabilité ?

Outre tous ces freins, la question la plus sensible repose certainement sur l'utilisateur. Le Zero Trust suppose ou inclut la formation et la sensibilisation de ces derniers pour réduire le risque de Phishing comme d'attaques par ingénierie sociale. Reste ensuite à placer le curseur de ce qui est acceptable dans le cadre des tâches quotidiennes. En d'autres mots, une répétition trop fréquente d'un code peut se révéler contre-productive et se traduire par contournement des outils de l'entreprise (e-mails personnels…), une diminution de la productivité, voire un arrêt du travail. Illustration parlante, un service d'urgence hospitalier a vocation à faire appel à des soignants d'autres services… qui n'auront pas accès aux dossiers patients. « *Les technologies biométriques tout comme le MFA sont incontournables dans le principe du Zero Trust ou Zero privilège* », a pondéré Eddy Sifflet de Check Point. Des moyens qui ne peuvent malgré tout être mis en place dans une partie des usages pour des raisons éthiques ou réglementaires. Reste donc à placer le curseur entre utilisabilité et contrôle avec des utilisateurs. Au final, si restreindre les droits sur une application, un serveur… aux objectifs supposés de l'utilisateur ou d'un appareil parait fondé, un Zero Trust trop poussé et rigide pourrait se révéler contre-productif en particulier parce qu'il est susceptible de générer un nombre important de faux positifs.

Enfin, l'habituelle dynamique du secteur numérique a tendance à survendre tout "concept" récent. Selon le site web de l'ANSSI, ce modèle attractif pour les entreprises et organisations, notamment parce qu'il est adapté au télétravail et au BYOD, fait « *l'objet d'un engouement de la part d'éditeurs de solutions technologiques et de sécurité qui y voient la perspective de nouveaux gains* ». Pas de vraie surprise, mais une raison de plus pour bien dimensionner les projets, bien placer le curseur entre sécurité et utilisabilité et, pourquoi pas, (re)mettre un peu plus d'utilisateur dans la boucle de la confiance. ▮

P.BR.

1 : https://federalnewsnetwork.com/commentary/2025/03/achieving-unified-visibility-for-effective-zero-trust-implementation/

> ***« Dans les grands comptes, le Zero Trust ne couvre pas encore la totalité des systèmes d'information. Il reste des bastions sur le modèle legacy »***
>
> Alexis Touret,
> expert en cybersécurité, Zscaler

« Avec le cumul des fonctions de DSI et de DPO, il est difficile de se contrôler soi-même »

Patrick Blum, délégué général de l'AFCDP, ex-DSI-DPO de l'Essec

Délégué général de l'AFCDP, ancien DSI et DPO de l'Essec, Patrick Blum pointe du doigt l'indépendance toute relative du délégué à la protection des données qui serait également DSI.

Etiez-vous un DSI sensible à la conformité ?

Patrick Blum : Lorsque j'étais DSI de l'Essec, j'étais attentif à la conformité et à la protection des données des personnes, dès 1983. En 2004, je suis devenu CIL, puis DPO avec l'arrivée du RGPD, en conservant mon rôle de DSI. Durant cette période, j'ai constaté que cette double responsabilité peut poser un problème d'indépendance et même de conflit d'intérêt.

Pouvez-vous préciser à quel niveau ?

PB : Le DSI décide de la mise en œuvre des traitements. La fonction de DPO comporte une mission de conseil auprès du responsable de traitements, mais aussi des missions de contrôle et d'audit. Avec le cumul des fonctions de DSI et de DPO, il est difficile de se contrôler soi-même. Cela demande une extrême rigueur, en pratique. Et, juridiquement, c'est discutable.

Deviez-vous faire des concessions parfois ?

PB : En pratique, je ne faisais pas de concession. Ma fonction de CIL, puis celle de DPO, prenait le dessus. A chaque nouveau traitement, je m'assurais que tout soit fait dans le respect des règles, sans prendre de demi-mesures. Oui, c'était compliqué. A présent, avec l'avènement de l'IA, les traitements posent des problèmes encore plus complexes au DPO. Ces technologies succèdent au CRM et à la vidéosurveillance qui posaient déjà des questions spécifiques.

Comment innover sereinement en IA ?

PB : Je constate souvent que certains traitements d'IA n'ont pas été signalés ni même aperçus par le DPO. Or, dès lors qu'ils utilisent des données personnelles, le DPO devrait s'assurer qu'ils respectent bien les droits des individus. Trop souvent, le RSSI et le DPO travaillent dans leur coin. Comme la sécurité fait plus peur, le RSSI est souvent le plus écouté des deux. Mais, en cas de fuite de données personnelles, le RSSI et le DPO sont concernés tous les deux. Ils doivent donc travailler ensemble.

Le responsable RSE est-il impliqué ?

PB : La sécurité du SI et sa conformité sont censés s'intégrer dans une responsabilité sociétale. Lorsqu'il y a un responsable RSE dans l'organisation, la logique voudrait qu'il soit en relation avec le RSSI et avec le DPO.

La chaîne logistique restant fragile, les sous-traitants participent-ils aux efforts de résilience ?

PB : Les mêmes risques existent pour une petite et une grande entreprise. D'autre part, la loi s'applique aux structures de toutes tailles. Une startup ou une société unipersonnelle peuvent avoir du mal à trouver les moyens d'assurer leur cybersécurité. Elle n'est pas tenue d'affecter le rôle de DPO à un salarié. Une solution est prévue dans le RGPD qui consiste à faire appel à un DPO externe. Il faut alors établir un contrat de services avec une personne ou une entreprise dotée de bonnes compétences en droit et en technologies.

Un patron de PME externalisant ses traitements auprès d'un hébergeur peut-il lui confier les rôles de DPO, ou de RSSI ?

PB : Ce ne serait pas une bonne idée. Si l'on fait appel à un hébergeur pour gérer une partie de son infrastructure, ce dernier va la mettre en œuvre, mais il ne pourra pas se contrôler en même temps, faute d'indépendance. Mieux vaut prévoir un contrat de mission avec un autre prestataire assurant la fonction de DPO indépendant, dans ce cas.

En cas de fuite d'un bulletin de paie sous-traité, qui est responsable ?

PB : Une petite structure sous-traitant sa paie auprès d'un prestataire n'a pas de maîtrise sur les traitements mais, comme c'est elle qui a décidé de faire appel aux traitements externalisés, elle doit bien vérifier la nature du contrat de sous-traitance et ses clauses. En cas de fuite de données à caractère personnel, la CNIL pourra contrôler les barrières prévues et les audits demandés afin de déterminer la responsabilité de chacun.

Faut-il encadrer les usages dans l'entreprise d'outils OSINT permettant de capter des renseignements sur les concurrents, les salariés, les prospects ou les clients ?

PB : Dès qu'un salarié de l'entreprise met en œuvre un traitement, y compris sous une feuille de calcul, cela entre dans la responsabilité de son employeur. Si jamais le salarié fait des choses illicites, le patron peut mettre en place des sanctions, mais il reste le responsable des traitements. S'agissant des traitements à base de recherches sur les données en source ouverte, c'est la même chose. D'où l'intérêt de ne pas négliger la rédaction d'une charte informatique, avec l'aide d'un juriste, pour préciser les usages permis et interdits.

Propos recueillis par Olivier Bouzereau

© Alice Fenwick

SMART TECH

DELPHINE SABATTIER

7H30 | 18H30

VOTRE ÉMISSION QUOTIDIENNE DÉDIÉE À L'INNOVATION

Dans l'émission SMART TECH animée par Delphine Sabattier,
l'actualité du numérique et de l'innovation prend tout son sens.
Chaque jour, des spécialistes décryptent les dernières news, les tendances,
et les enjeux soulevés par l'adoption des nouvelles technologies.

N°230 orange™ | N°246 bouygues | N°163 free

B SMART by I. Change

TOPTECH

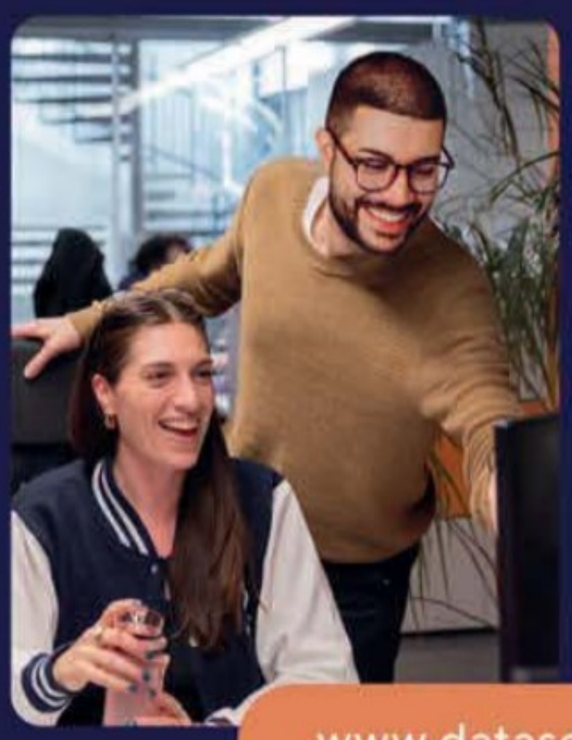
Un projet ? Contactez-nous • Un projet ? Contactez-nous •

DATASOLUTION

YOUR DIGITAL FACTORY

DEPUIS **Concepteur** 2003

d'expériences digitales

www.datasolution.fr

E-commerce & CMS

Marketplaces, CMS, Commerce Unifié

Référentiel de données

MDM, PIM, DAM, ERP, RCU

Digital Services

Studio Créa, RGPD, Business Performance

Plateforme production marketing

Web-to-print, Solution de personnalisation, Production de packaging

Cloud Services

Cloud public & privé, Services managés, IA, Cybersécurité