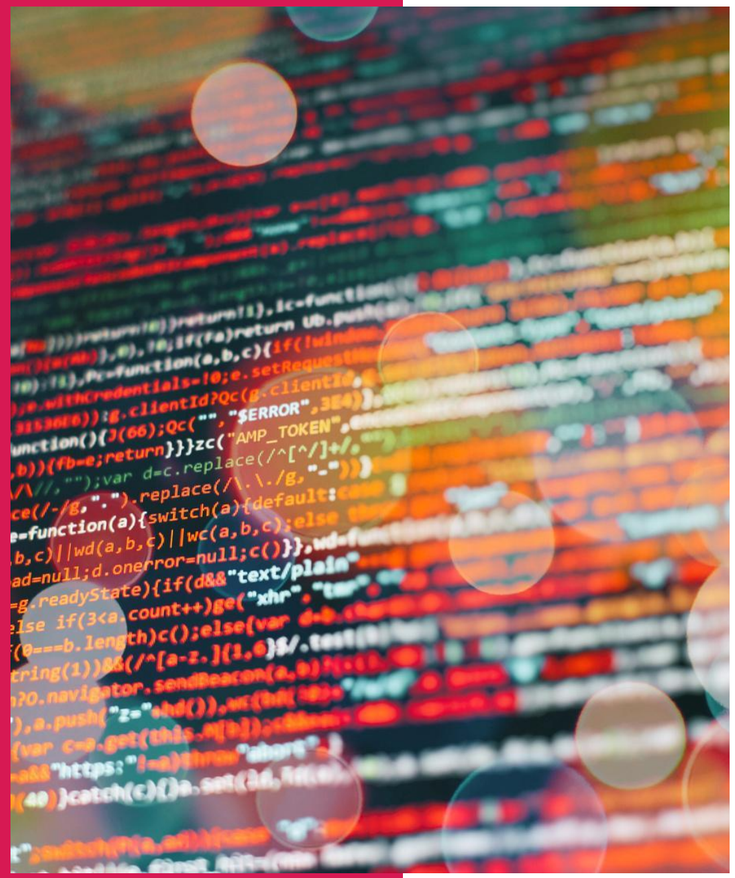


TENDANCES TECH

[sfeir]
WENVISION

#TechTrends

Au menu de 2024 : TOC, MACH, RAG et CUI

4 principes directeurs pour investir dans un SI compatible avec le futur

Comme si l'industrie informatique ne regorgeait pas de sigles et acronymes, il fallait que nous en fassions des étendards de ces TechTrends 2024...

On vous l'accorde, les sonorités ne sont pas tellement sexy. Mais les concepts sous-jacents, en revanche, sont à la fois simples, séduisants et extrêmement puissants. Pris tous ensemble, ces concepts peuvent servir de principes directeurs pour guider les investissements de cette année dans un SI compatible avec le futur.

- **CUI, pour Conversational User Interface**, prend acte du succès du chat dans ChatGPT. Aujourd'hui, on utilise un système d'information. Demain, on collaborera avec le SI, au travers d'un dialogue en langage naturel.
- **RAG, pour Retrieval Augmented Generation**, est LA technologie qui permet à l'IA générative d'exprimer toute sa valeur en entreprise. Avec du RAG, les données ou documents de l'entreprise deviennent accessibles aux IA génératives, de manière à apporter des réponses bien plus pertinentes.
- **MACH** est déjà un acronyme, regroupant 4 éléments à activer ensemble pour en tirer tous les bénéfices : des **microservices**, ou a minima une architecture modulaire, des **API** pour faciliter le dialogue dans un système découplé, une plate-forme conçue nativement pour le **Cloud**, et des applications détachées de leur front (**headless**) de manière à pouvoir recomposer et réorchestrer les services selon les besoins.
- **TOC**, enfin, presque le plus important : parce que la plus belle **Tech** du monde ne donnera pas les résultats escomptés si l'**Orga** et les processus ne sont pas adaptés, ce qui implique une **Culture** d'entreprise ouverte à ces changements et des collaborateurs qui auront pu s'acculturer et se former.

Bienvenue dans ces TechTrends, bonne lecture !

SOMMAIRE

01. TECH
#GenAI
#Cloud

02. ORGA
#Equipes
#CareerPaths

03. CULTURE
#GrandsPrincipes
#Pilotage&Ethique

04. PLATEFORMES
#Archi&Dev
#SICloud

05. PRODUITS
#ProcessusProduits
#ProductTrends

06. DATA
#DataConcepts
#DataGouv

TECH

/01

TECH

LES CONTRIBUTEURS DE CETTE SECTION

ILS ET ELLES ONT ÉCRIT LES TECH TRENDS 2024

François DENIS
Head of Cloud Advisory,
WENVISION

Ravidhu DISSANAYAKE
Consultant Data / IA,
WENVISION

Martin ELIARD
Consultant Data / IA,
WENVISION

Marie FONTAINE
Head of AI
WENVISION

Didier GIRARD
Co-CEO,
SFEIR & WENVISION

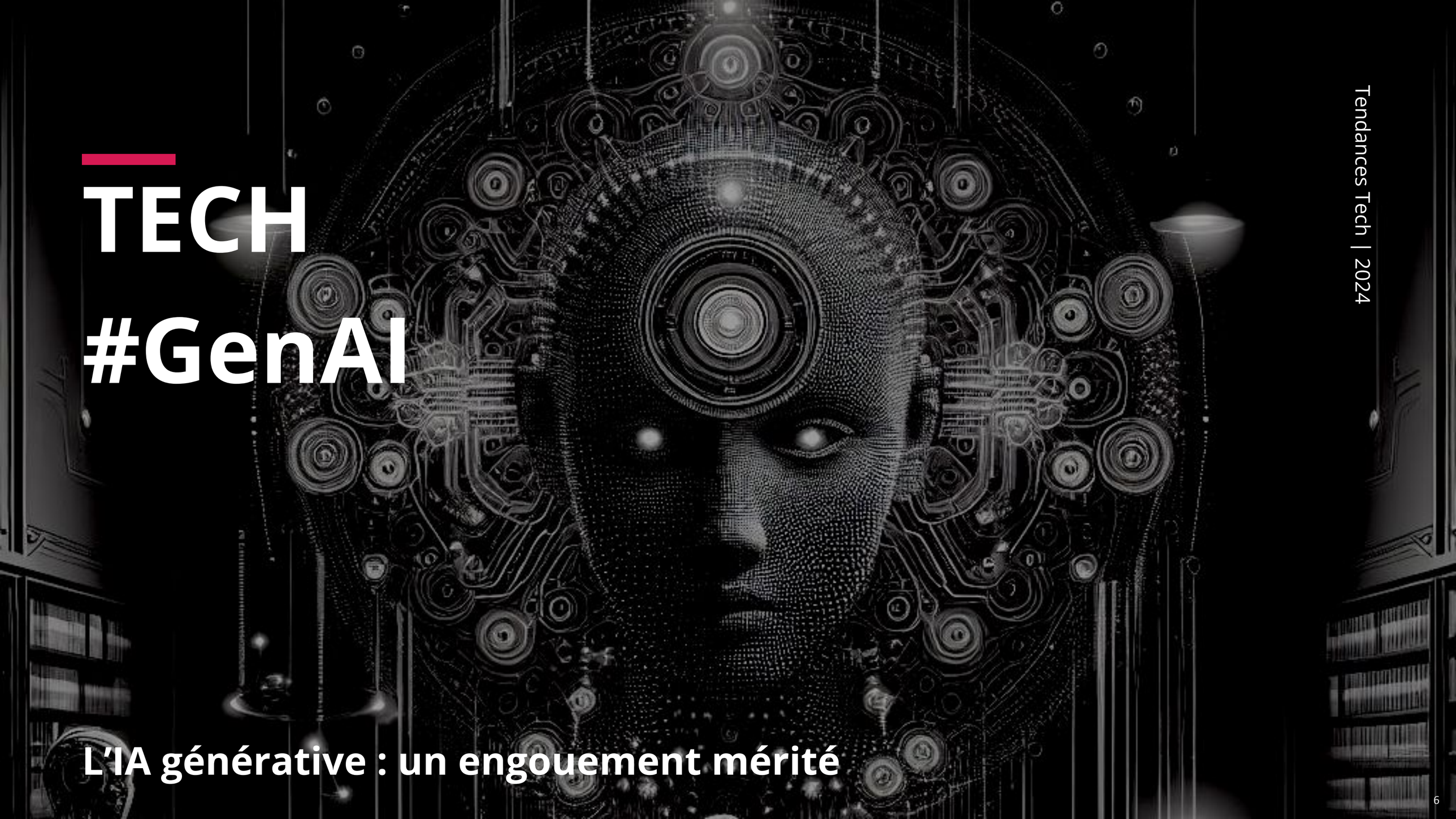
Mylène MAIGNANT
Consultante IA,
WENVISION

Seifeddin MANSRI
CTO Cloud Groupe,
SFEIR

Baptiste PUGNAIRE
Consultant Data / IA,
WENVISION

Olivier RAFAL
Consulting Director Strategy,
WENVISION

TECH
#GenAI



L’IA générative : un engouement mérité

#Tech #GenAI

L’IA générative est parmi nous - de manière durable

L'IA générative (Generative AI) est un type d'intelligence artificielle qui se concentre sur la génération de nouvelles données similaires à une entrée donnée. Des applications de génération d’images comme Dall-E 2, Stable Diffusion et Midjourney, ou de génération de texte ou de code comme GPT / ChatGPT ou encore GitHub CoPilot ou AWS Code Whisperer, ont très rapidement mis ces capacités sur le devant de la scène.

3 techniques clés

Une des techniques clés utilisées en IA générative est l'utilisation de réseaux génératifs adverses (generative adversarial networks / GAN), qui se composent de deux composants : un générateur et un discriminateur. Le générateur produit des données synthétiques basées sur l'entrée, tandis que le discriminateur tente de déterminer si les données sont réelles ou fausses. Cette compétition entre les deux réseaux permet à l'IA de générer des données de plus en plus similaires à l'entrée.

Les auto-encodeurs à variation (Variational autoencoders / VAE) sont une autre technique importante utilisée en IA générative. Ces algorithmes se composent de deux éléments : un encodeur, qui comprime les données d'entrée en une représentation à dimension inférieure, et un décodeur, qui prend cette représentation et génère de nouvelles données similaires à l'entrée.

Les transformateurs (transformers) sont une troisième technique importante en IA générative. Ces algorithmes utilisent un mécanisme dit d'attention pour mesurer et prédire l'importance des différentes parties des données d'entrée. Les transformers sont à la base des progrès faits récemment dans le domaine du traitement automatique du texte (Le T de BERT / GPT).

Une infinité d’applications et de cas d’usage

L'IA générative est un domaine en évolution rapide qui suscite à juste titre énormément d’intérêt en raison de ses applications potentielles dans des domaines tels que l'augmentation des données, la génération d'images et le traitement du langage naturel. Cette technologie a de nombreuses applications potentielles, typiquement :

- Juridique : l'IA générative peut générer rapidement des contrats et autres documents juridiques, ou identifier des modèles dans des affaires juridiques qui pourraient aider à prédire les résultats de futurs cas ou a minima à construire des dossiers.
- Média : l'IA générative peut aider à créer de nouvelles images, vidéos, présentations ou autres formes de contenu à partir d'une entrée donnée, comme une photographie ou une description écrite, où à la manière d’un artiste célèbre (la banque d’images Shutterstock a déjà conclu un accord pour commercialiser des images issues de Dall-E 2).
- Éducation : l'IA générative peut créer des plans de leçon personnalisés pour les élèves, ou générer automatiquement des questions de quiz et d'autres matériaux pédagogiques.
- Développement de logiciels : l'IA générative sait automatiser la création de code (pour interroger une API par ex.), ou aider à identifier les potentiels bugs et autres problèmes dans les logiciels.
- Assurance : l'IA générative peut contribuer à générer rapidement des polices d'assurance, ou identifier des modèles dans les données de réclamation qui pourraient aider à prédire les risques futurs.
- Banque : l'IA générative peut automatiser la création de rapports financiers, ou aider à écrire des messages commerciaux personnalisés.

Ce ne sont que quelques exemples des nombreuses applications potentielles de l'IA générative dans différents domaines. On voit d’ailleurs fleurir de nombreux projets d’apps ou de start-up. À mesure que la technologie continue d'évoluer, il est probable que de nouvelles utilisations innovantes seront découvertes.

#Tech #GenAI #UseCases

Les tendances d’usage de l’IA Générative en 2024

Les usages principaux de l’IA Générative peuvent être organisés autour de 2 grandes catégories : **la recherche et la création de contenus.** Partant de là, prétendre pouvoir lister les cas d’usage les plus pertinents pour les entreprises en 2024 relève de l’oracle tant l’horizon semble ouvert et tant la “bonne” utilisation d’une technologie repose sur les priorités stratégiques de l’entreprise. On peut néanmoins noter que la tendance est plutôt aux usages internes, moins risqués, maximisant plutôt les gains de productivité (assistants généralistes, assistants verticaux, assistants de programmation, générateur de contenu pour le marketing, la communication, les centres de relation client).

Les capacités en entreprise pour répondre aux besoins sont limités, nous observons donc plutôt un besoin de gouverner ces usages avec deux questions centrales : comment organiser leur collecte et surtout comment prioriser parmi ce champ des possibles ? L’application des principes du Product Management, avec l’ajout de quelques spécificités, répond à ces questions.

Collecter les besoins par l’usage d’une solution d’assistant interne

S’il ne fallait choisir qu’un usage pour 2024, celui de l’assistant généraliste pour les collaborateurs semble le plus intéressant. Il permet de lever des freins de productivité pour de nombreuses fonctions et il peut surtout servir de première marche pour construire un socle technologique, des compétences et collecter les besoins des collaborateurs par l’usage concret qui en est fait. Cela permettra par la suite de verticaliser des besoins à travers des assistants spécialisés par exemple. C’est le chemin que plusieurs de nos clients ont entrepris en 2023.

La priorisation des usages doit suivre les fondamentaux du product management

En premier lieu, il est absolument nécessaire de valider la réponse à un besoin et de vérifier l’adéquation des hypothèses de gain avec les objectifs stratégiques de l’entreprise. Cette première étape doit également d’ores et déjà inclure une vérification de la conformité de la finalité de l’usage au regard de la réglementation, à savoir l’AI Act.

Un facteur de risque peut aussi être appliqué en fonction de l’acceptabilité d’une réponse fausse produite par le système. C’est notamment pour cette raison que **les systèmes d’IA générative mis directement à disposition des clients seront rarement priorisés en 2024, au profit d’usages tournés vers l’interne** avec des hypothèses de gain de productivité.

La complexité doit être mesurée, avec quelques spécificités dont il faut tenir compte : l’accessibilité de la donnée, la maturité des briques technologiques en jeu, le degré de sécurité attendu.

Un framework tel que RICE (Reach, Impact, Confidence, Effort) adapté à l’IA Générative peut s’avérer un outil de gouvernance très efficace.

#Tech #GenAI #LMM

Du LLM aux LMM - les modèles multimodaux

Le multimodal : l'IA générative pour l'analyse des documents courants

Les LLM sont spécialisés dans la donnée textuelle. Ils rencontrent donc des limites lorsqu'il s'agit d'interpréter des données multimodales, qui comprennent non seulement du texte mais aussi des images, des sons et des vidéos. L'avenir de IA générative réside ainsi dans l'utilisation de modèles multimodaux.

Les documents d'entreprise contiennent souvent des éléments visuels comme des images, des graphiques et des schémas. Ces éléments sont riches en informations contextuelles essentielles à la compréhension du document. Les modèles multimodaux, dits LMM, permettent de combler le fossé entre le traitement du langage naturel et la perception visuelle.

En combinant la puissance des LLM avec la capacité de traiter des données visuelles, ces modèles promettent :

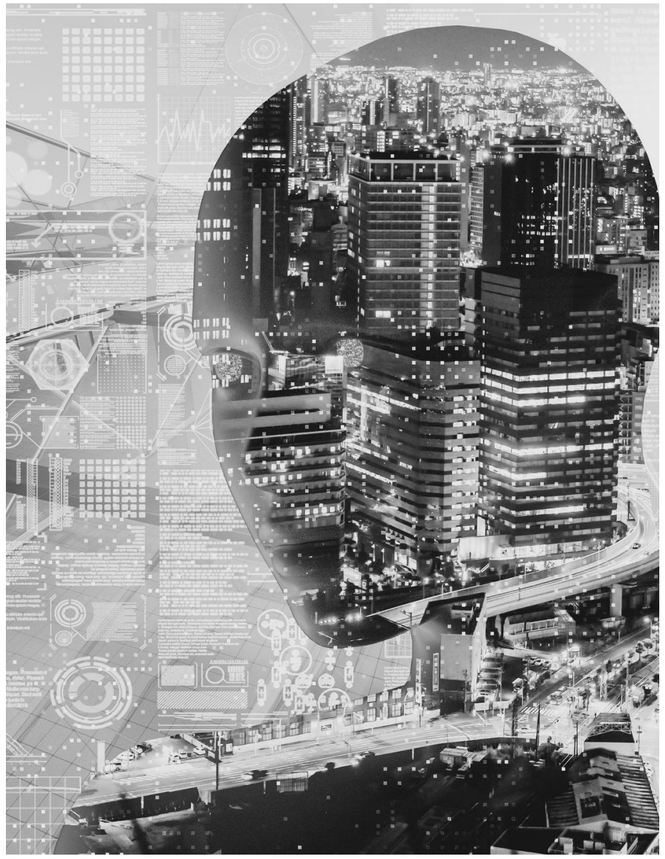
- Une extraction d'informations plus efficace : les modèles multimodaux peuvent extraire automatiquement des informations à partir d'images et de textes, ce qui rationalise la gestion des documents et rend les informations plus accessibles.
- Des générations plus riches : en intégrant des données visuelles, les modèles multimodaux peuvent générer des textes plus complets.
- Des recommandations plus personnalisées : dans le domaine du commerce électronique, ces modèles peuvent fournir des recommandations personnalisées en tenant compte des images, des descriptions de produits et des évaluations des utilisateurs.

Exemples d'applications multimodales améliorées par les LMM : rédaction automatique de résumés de documents, recherche d'information multimodale, assistance à la clientèle à partir de photos et de texte, etc.

Les applications GenAI multimodales seront dans les roadmaps 2024 des entreprises

#Tech #GenAI #SLM

LLM ou SLM ? Les modèles de langage se mettent au régime



Travailler avec des Large Language Models implique des contraintes notamment de coûts et d'infrastructure. Les Small Language Models (SLM) se démarquent de leurs homologues plus imposants en offrant une approche plus efficiente, flexible et personnalisable. Pour créer un SLM, on a d'abord besoin d'un LLM pour servir de référence (ou de fondation) et, par un processus de "distillation", on parvient à entraîner un modèle plus compact.

L'objectif poursuivi lors de la création d'un SLM est d'optimiser un rapport bénéfices/coûts : on cherche à maintenir un niveau de performance comparable au LLM de référence en réduisant progressivement le nombre de paramètres nécessaires.

Une fois entrainés, ces SLM sont beaucoup plus flexibles car ils nécessitent moins de puissance de calcul à l'usage et occupent moins de place dans la mémoire des appareils qui l'embarquent. Cette approche ouvre de nouvelles possibilités pour décentraliser l'utilisation des modèles de langage en permettant leur déploiement dans des systèmes du quotidien comme nos ordinateurs et téléphones portables. D'autre part, la réduction du nombre de calculs nécessaires pour effectuer des prédictions fait des SLM des bons candidats pour les applications nécessitant du temps réel ou une forte réactivité comme la retranscription vocale.

Des modèles spécialisés, experts dans un domaine

Il est particulièrement difficile de personnaliser un LLM : cette opération peut être instable, nécessite énormément de données de qualité, très coûteuse également... C'est principalement pour cela que si peu d'entreprise sont capables de proposer de réelles alternatives à ChatGPT. Les SLM permettent de faire sauter cette barrière à l'entrée en diminuant toutes ces complexités (proportionnellement à la réduction du nombre de paramètres). On peut donc plus facilement obtenir des modèles spécialisés qui, en moyenne, sont moins bons qu'un modèle généraliste, mais font preuve d'expertise dans leur domaine (un peu comme nous finalement).

Un bénéfice moins connu de cette spécialisation est la diminution des hallucinations : en réduisant le périmètre de compétence, on facilite le choix d'un corpus d'entraînement très pertinent qui limite aussi l'exposition de données sensibles.

#Tech #GenAI #LAM

Le début des agents autonomes

Rabbit r1, pionnier des “assistants de poche”, dopé au LLM Perplexity

Dans la famille des déclinaisons des LLM, voici venir les LAM : Large Action Models. L’idée est d’aller au-delà d’assistants “passifs”, qui répondent à des requêtes, pour créer de véritables agents “actifs”, opérant des actions.

Il n’est toujours pas question de remplacer l’humain, mais bien de l’augmenter, en lui fournissant un assistant capable de réaliser à sa place une suite de tâches répétitives, sans valeur ajoutée ; c’est en quelque sorte une revisite de la promesse du RDA (Robotic Desktop Automation) rendue possible par l’IA générative.

Le rôle des LAM : orchestrer et automatiser les assistants LLM

Le RDA automatisait une suite de tâches précises (ouvrir telle application, créer un utilisateur, noter son ID, ouvrir une 2e application, y entrer cet ID, etc.) à partir d’un input prédéfini. Grâce aux capacités de traitement du langage naturel, l’objectif des LAM est de permettre à l’utilisateur de définir un objectif final et de laisser l’agent déterminer l’enchaînement des actions à accomplir pour y parvenir.

Le tout récent Rabbit r1 en donne un premier exemple (perfectible) dans la sphère personnelle : son ambition est de remplacer le besoin d’interagir manuellement avec les apps d’un smartphone par des commandes vocales en langage naturel (par exemple, réserve-moi un VTC pour aller à telle heure à tel endroit).

Dans le monde professionnel, le LAM viendra automatiser et orchestrer des actions impliquant plusieurs assistants LLM. La priorité est donc à l’industrialisation des assistants, mais il faut avoir en tête cette évolution !

#Tech #GenAI #RAG

Vectoriser les données de l’entreprise pour dialoguer avec le SI

Agacé par un résultat non pertinent de votre IA générative préférée, vous avez déjà tenté de changer des éléments de votre prompt ? Si oui, alors comme M. Jourdain qui pratiquait la prose naturellement, vous avez eu recours au prompt engineering.

Le résultat fut-il meilleur ? Probablement, mais certainement toujours loin de vos attentes initiales. Vous avez ainsi touché du doigt les limites de "l'ingénierie de la requête". Le prompt engineering est une vraie compétence, utile, mais il serait illusoire de penser 1) les utilisateurs maîtriseront cet art, et 2) que cela résoudra les soucis de pertinence du contenu.

Le vrai sujet : offrir un dialogue avec le SI en langage naturel

Ce qui donnera véritablement de la valeur à vos applications d'IA générative en entreprise, c'est la capacité offerte aux collaborateurs d'accéder au patrimoine informationnel à travers un dialogue en langage naturel. D'obtenir des réponses pertinentes pour leur domaine d'activité, pour leur métier. Le niveau d'optimisation de la requête sera un plus, qui pourra s'affiner au fil du temps.

Cela nécessite de mettre en place du RAG (Retrieval Augmented Generation) : une architecture qui va combiner la puissance des LLM pour la compréhension des questions et la formulation des réponses et le contenu des documents des utilisateurs (RAG sur des PDF individuels) voire des bases d’informations des entreprises.

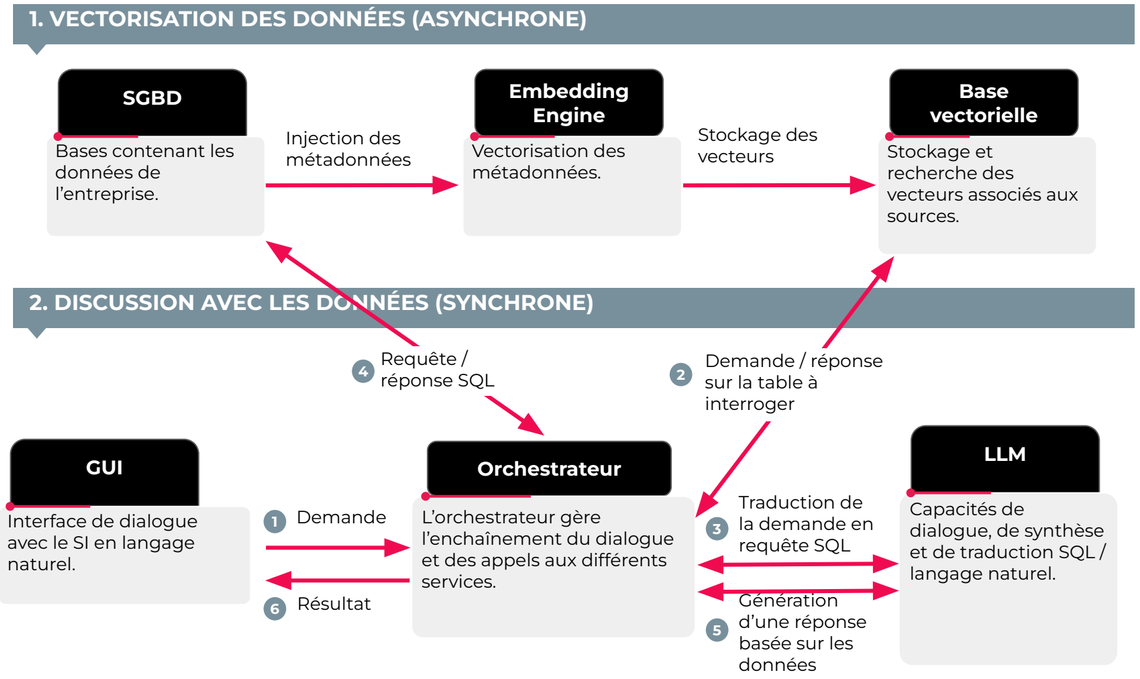
Donner ainsi accès aux données et documents demande de les vectoriser au préalable. Typiquement, vous disposez de dizaines de documents techniques, des PDF de plusieurs dizaines de pages, et vous souhaitez permettre à vos collaborateurs d'interroger cette base de connaissance en langage naturel comme s'ils discutaient avec un documentaliste. Il faut alors passer par une phase dite d'embedding : la création de vecteurs représentant leur contenu, que les moteurs d'IA générative sont capables d'interpréter.

Embedding de documents et données structurées

Cette vectorisation du contenu de l'entreprise est aussi réalisable pour les données structurées : chiffres de vente, nombre d'employés, indicateurs RSE, etc. C'est là où la promesse de l'IA générative en entreprise devient exceptionnelle. C'est là également que le prompt engineering prend tout son sens : selon l'habileté qu'on aura à formuler des prompts pour dialoguer avec le SI, les résultats pourront être très différents.

D'où l'importance de coupler ces nouvelles interfaces de chat avec des bibliothèques de prompt : des questions peaufinées, dont on sait qu'elles généreront des résultats pertinents.

Bien entendu, cela demande de prendre quelques précautions. Les bases de données doivent être correctement renseignées et les métadonnées intelligibles. Il faut trouver le bon découpage des documents (en “chunks” analysables par les LLM). Il faut aussi se prémunir contre des risques de sécurité (SQL injection). Mais très clairement, cette approche de requêtage en langage naturel est idéale pour de très nombreux utilisateurs qui pourront s’appuyer sur un assistant personnel pour naviguer au sein du patrimoine informationnel de l’entreprise.



#Tech #GenAI #Gouvernance

Des IA génératives d’entreprise, pour limiter risques et shadow AI

Interdiction, blocage, bannissement : la presse regorge d'exemples d'entreprises ayant interdit ChatGPT. Si Samsung, Apple, Deutsche Bank, Verizon ou Amazon l'ont fait, c'est qu’il y a une raison : éviter la fuite de données. Il faut considérer les IA génératives grand public comme des photocopieuses en libre accès, qui peuvent à tout moment ressortir vos documents, briefs stratégiques, etc. Elles doivent donc être interdites. Ce qui veut dire que l’entreprise doit fournir un service équivalent dans un cadre sécurisé.

D’une part, parce qu’il serait idiot de se priver du potentiel de l’IA générative, mais aussi de monitorer l’usage qui en est fait. D’autre part, parce que les utilisateurs en entreprise trouvent toujours un moyen de contourner les interdictions lorsqu'ils trouvent quelque chose de beaucoup plus pratique. En l'occurrence, si l'usage de ChatGPT et des autres applications d'IA générative a explosé aussi rapidement, c'est bien parce que ces outils apportent une aide énorme dans le quotidien. Ne pas offrir de solution, c’est s’exposer à l’ émergence rapide d’un “shadow AI”, susceptible de mettre l’entreprise à risque.

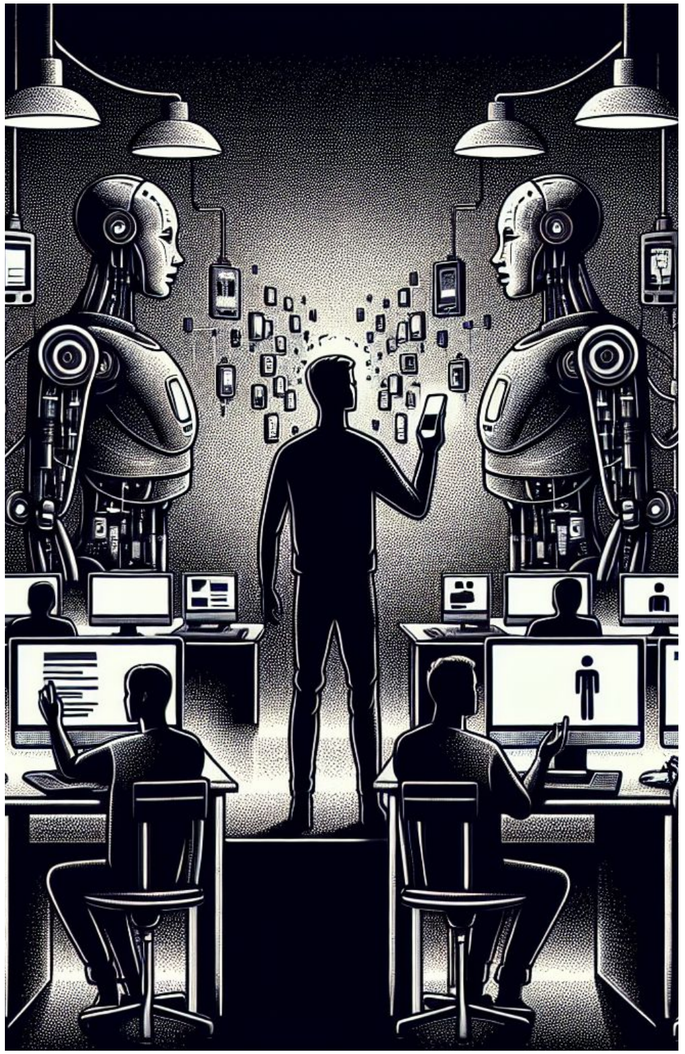
Une stratégie mixant “make” et “buy”

Plusieurs pistes sont envisageables pour mettre à disposition des environnements professionnels, sécurisés. Aucune ne consiste à développer son propre modèle. Il n’est généralement pas non plus pertinent de le “fine-tuner” pour optimiser son usage : c’est un processus lent et coûteux, à réserver à quelques cas précis.

Partir sur un environnement tout-en-un est possible mais risqué, car on s’en remet alors à un éditeur et sa roadmap, sans possibilité de supervision. L’approche pragmatique que nous préconisons et voyons plutôt se développer dans les entreprises consiste à mixer “make” et “buy”. Autrement dit, s’appuyer sur les ressources des fournisseurs comme Google, Microsoft, Amazon, OpenAI, HuggingFace, Mistral, etc. et consacrer ses efforts à l'assemblage de ces ressources - par exemple une base de données vectorielle pour introduire de la sémantique au sein de ses propres documents, un orchestrateur type Langchain pour enchaîner les appels à des services ou encore des studios pour créer, optimiser et versionner les prompts utilisés.

Expliquer, acculturer

En plus d’interdire et de fournir une solution, il est impératif de communiquer et d’acculturer les collaborateurs à cette technologie et à ses enjeux. Il s’agit de faire comprendre pourquoi les outils grand public sont interdits, de montrer les forces ainsi que les faiblesses et les risques associés à la GenAI (biais, hallucinations, etc.), mais aussi de donner des perspectives sur les futures possibilités techniques qui seront mises à disposition et les cas d’usage qui pourront ainsi être mis en œuvre. Ces réflexions doivent s’inscrire dans un cadrage stratégique, qui lui-même doit respecter la vision et les valeurs de chaque entreprise.



#Tech #GenAI #ChatUI

L’IA révolutionne nos interactions avec le SI

L'interaction avec un SI devrait être aussi simple que de converser avec un collègue. Et pourtant, depuis l'arrivée des "menus" dans les applications, nous ne cessons de cacher l'information en la rendant accessible uniquement après une suite d'interactions. L'objectif d'une interface conversationnelle est de simplifier et de rendre plus intuitive l'accès aux données.

Il n’y a qu’à observer le succès quasi-instantané de ChatGPT pour comprendre le potentiel de ce mode d’interaction. Alors que la technologie existait depuis plusieurs années, l’ouverture d’un accès en mode conversationnel a fait exploser son usage.

Les CUI (conversational user interfaces) ne sont pas seulement intuitives par nature, avec un besoin en formation minimale, elles éliminent également le besoin de naviguer à travers des menus hiérarchiques complexes et chronophages. En s'adaptant au niveau d'expertise de l'utilisateur, les CUI offrent une expérience personnalisée. L'expert peut poser ses questions d'expert, l'utilisateur occasionnel peut lui aussi interagir en posant ses propres questions.

Des CUI (conversational user interfaces) en complément des traditionnelles GUI

C'est la technologie qui se plie à la volonté de l'utilisateur : elle comprend le langage "parlé", et transforme la façon d'interagir avec les données, l'analytique et le prédictif.

L'avantage est l'accélération des tâches routinières par l'action de simples commandes. Que l'on consulte des bases de données, génère des rapports ou analyse des tendances, les CUI offrent une efficacité que les GUI (interfaces graphiques, graphical user interfaces) traditionnelles ne peuvent égaler.

Toutefois, il ne faut pas enterrer trop vite nos bonnes vieilles GUI ; elles ont toujours leur place. Pour les tâches complexes et spécifiques au sein du SI, la précision de la GUI classique reste. Ce parfait équilibre assure que, tandis que les CUI gèreront la majorité des interactions en améliorant la productivité, les GUI resteront disponibles pour gérer les interactions plus fines.

#Tech #GenAI #Challenges

IA générative : 3 grands challenges sociaux en perspective

Les *fake news*

Alors que les technologies d'IA continuent de se développer à un rythme effréné, elles sont devenues à la fois une arme et un rempart dans la propagation et la détection des fausses informations. Le défi réside dans la création de modèles d'IA suffisamment intelligents pour discerner les faux des vrais faits (textes, images, vidéos…), tout en garantissant la protection de la liberté d'expression et en évitant les biais préjudiciables. Une gageure.

En 2024, les populations de 77 pays, dont les Etats-Unis, la Russie et l'Inde, seront amenées à voter. Dans ce contexte électoral mondial, la gestion des *faux* voire des *vrais faux* prend une dimension encore plus critique. Exemple frappant : au Pakistan, en décembre 2023, un parti politique a généré un discours de son leader, emprisonné, à partir de vidéos archivées, afin de pouvoir tout de même mener la campagne des législatives.

La consommation énergétique

La demande croissante en serveurs de données, en centres de calcul et en infrastructures de refroidissement liée à l'IA générative contribue à l'augmentation de l'empreinte carbone du numérique. Les enjeux de la consommation d'énergie par l'IA en 2024 sont doubles : d'une part, il s'agit de réduire cette empreinte en développant des technologies et des infrastructures plus éco-énergétiques. D'autre part, il est primordial de sensibiliser et d'inciter les acteurs de l'industrie ainsi que les utilisateurs finaux à adopter des bons réflexes et des pratiques plus durables.

L'AI Act européen

L'AI Act en cours de discussion au sein de l'Union européenne illustre la prise de conscience croissante des enjeux liés à l'IA, notamment en matière de protection des droits fondamentaux et de prévention de la discrimination. Toutefois, sa future application soulève également des défis, notamment en ce qui concerne la mise en conformité des entreprises et l'équilibrage entre l'innovation et la réglementation. Même s'il est difficile de prévoir à quelle échéance ce texte sera en vigueur, il est clair que d'ores et déjà, la régulation de l'IA s'annonce comme un défi majeur à prendre en considération dès cette année.

#Tech #GENAI #Challenges

IA générative : 4 grands challenges techniques à relever

Conférence scientifique de référence sur l'IA, NeurIPS a dévoilé en décembre 2023 des avancées et des tendances clés pour 2024. **Pour nous, quatre thèmes majeurs façonneront de manière significative l'année à venir.**

Des résolutions de problèmes plus complexes

La technique "Tree of Thoughts" a émergé comme une avancée majeure, permettant une structuration plus fine du processus de pensée. Evolution du modèle actuel de "chain of thoughts", cette nouvelle technique consiste à structurer le processus de pensée du modèle sous forme d'arbre. Chaque branche de cet arbre représente une étape intermédiaire ou une sous-question qui aide à décomposer un problème complexe en éléments plus gérables. Ainsi, au lieu d'aborder un problème dans son ensemble, le modèle suit un chemin logique, examinant différentes possibilités et solutions étape par étape, ce qui permet une analyse plus approfondie et structurée.

Cette approche marque une progression vers des systèmes d'IA plus autonomes, capables de réflexion approfondie.

Des nouvelles architectures de LLM comme Mamba

La plupart des modèles LLM actuels sont fondés sur des Transformers (le T de GPT). Cette architecture a pourtant des limites en termes de longueur de séquence et de coûts d'exploitation. De nouvelles architectures prometteuses sont apparues pour y remédier.

Mamba en particulier révolutionne le traitement des longues séquences de données et offre des performances accrues. Cette architecture ouvre de nouvelles perspectives pour des applications exigeant le traitement de grandes quantités de données. On évoque des séquences pouvant aller jusqu'à 1 million de tokens !

Des IA autonomes et adaptatives

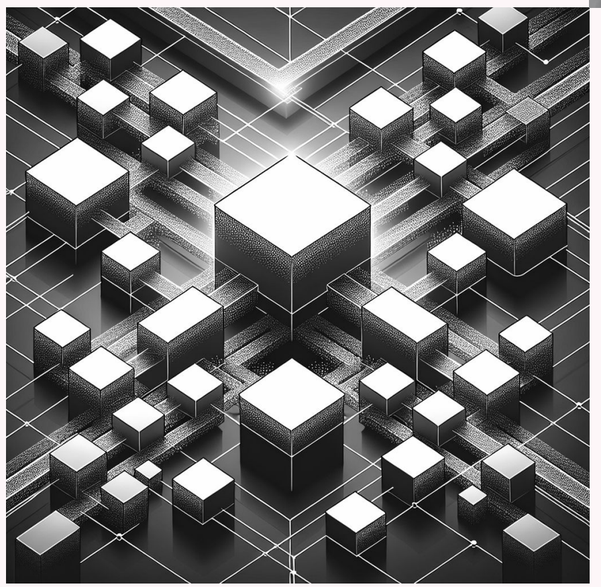
De nouveaux modèles et outils tels que Toolformer et Voyager soulignent la tendance vers des systèmes d'IA plus autonomes. Toolformer apprend à utiliser des outils externes, tandis que Voyager explore et acquiert des compétences dans des environnements virtuels, démontrant une capacité d'adaptation et d'interaction enrichie.

Des méthodologies d'évaluation des capacités des modèles

Les récentes études sur les LLM mettent en lumière la nécessité d'une analyse plus rigoureuse de leurs capacités cognitives. En se concentrant sur la planification et les cartes cognitives, ces recherches révèlent des lacunes notables dans la compréhension conceptuelle des modèles avancés tel que GPT-4. Les auteurs appellent à la création d'une méthodologie d'évaluation réaliste, soulignant que le chemin vers une IA authentiquement intelligente est parsemé d'obstacles méthodologiques et conceptuels.

#Tech #GenAI #Benchmarks

Les benchmarks LLM à suivre en 2024



Les benchmarks de LLM sont des jeux de données librement accessibles au public, conçus pour évaluer les performances et les capacités de ces modèles. Ces benchmarks sont créés par des chercheurs ou par des organisations et sont disponibles pour quiconque souhaite évaluer ses propres LLM. Ils ont 4 finalités :

- **L'évaluation des performances**

  Ils fournissent une méthodologie standardisée pour mesurer à quel point les LLM se comportent efficacement dans une variété de tâches, telles que la réponse à des questions, la création de résumés de texte, la traduction, etc.

- **La comparaison des modèles**

  Ils permettent aux chercheurs et aux développeurs de comparer différents LLM entre eux, en identifiant leurs points forts et leurs faiblesses.

- **Le suivi de la progression**

  Ils servent à suivre l'évolution des performances des LLM et à mettre en évidence les domaines qui nécessitent des améliorations.

- **Assurer l'équité et la transparence**

  En mettant à disposition du public ces benchmarks, d'autres utilisateurs peuvent ainsi reproduire et valider les résultats.

Il existe plusieurs manières d'évaluer des LLM : évaluations sans exemple préalable (on parle de "zero-shot learning") ou avec exemples ("few-shot learning"), celles fondées sur l'interaction humaine, etc., ainsi que différents critères selon lesquels leur évaluation est faite (créativité, interactions avec d'autres modèles, compréhension du comportement humain, etc.).

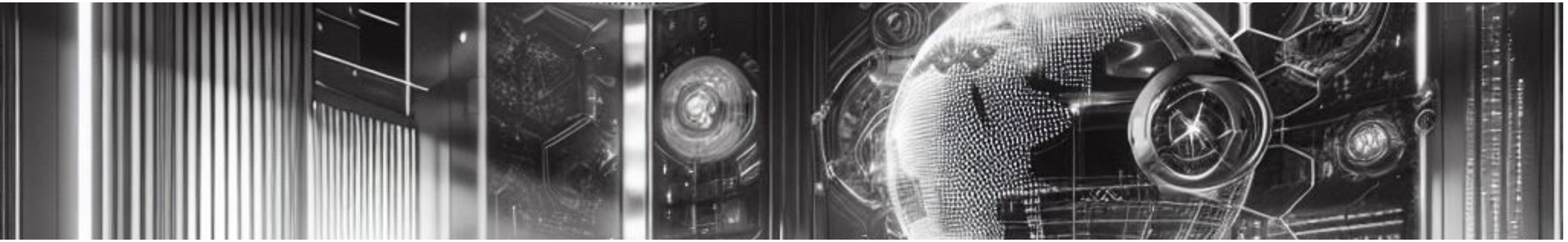
Open LLM Leaderboard offre un large panel intéressant de benchmarks. TriviaQA, HumanEval, GSM8K et MMLU restent autrement des ressources pertinentes et sûres pour aborder la question de l'efficacité des LLM.

TECH #CLOUD

Le Cloud reste un levier de transfo majeur

#Tech #Cloud

Le cloud permet de consacrer ses forces aux produits innovants



Le cloud est un outil puissant pour les entreprises, qu'elles soient petites ou grandes. Le cloud computing est essentiel pour l'innovation et accélère l'ensemble du processus de développement. Depuis son apparition, le cloud a révolutionné le fonctionnement des entreprises. Les démarches de type lift & shift sont désormais minoritaires : il s'agit de rapidement bénéficier de l'effet de levier du cloud, en adoptant les principes technologiques et en se réorganisant en conséquence.

Cloud first, cloud only, cloud native

Une stratégie "cloud first" est désormais adoptée par les entreprises de toutes tailles pour profiter des progrès rapides permis par le cloud computing. Cette stratégie amène rapidement à adopter le "cloud only". Et une fois à maturité, une entreprise peut se déclarer "cloud native".

Les entreprises qui adoptent cette philosophie peuvent consacrer leurs forces à la création de produits à valeur ajoutée, puisque les problématiques de plateforme technologique, de sécurité, de résilience, d'efficacité énergétique et de déploiement sur le territoire et au-delà sont déjà en très grande partie prises en compte par les fournisseurs de cloud.

Analyser les coûts et les risques, prendre les mesures appropriées

Le déploiement du cloud ne va pas sans peurs, qui pour certaines sont légitimes. Il est donc nécessaire pour les entreprises d'analyser froidement les risques et les coûts associés à cette technologie. Ces risques et ces coûts ne doivent pas stopper l'adoption du cloud mais permettre d'en sécuriser le déploiement, en adoptant les politiques et les règles idoines.

#Tech #Cloud #securite

L'essor des solutions de sécurité basées sur le cloud

Le cloud change la façon dont les entreprises opèrent. Il offre l'évolutivité, la flexibilité et l'efficacité d'un système d'information dont l'épine dorsale est Internet. À mesure que les organisations adoptent des solutions basées sur le cloud, l'importance de la sécurité de leur infrastructure cloud augmente pour s'adapter à ce changement de paradigme.

Les mesures de sécurité traditionnelles "on-premises" sont souvent non-adaptées pour faire face à ces nouveaux défis, créant un besoin critique de solutions de sécurité basées sur le cloud ou dites 'cloud native'.

Les solutions de sécurité cloud sont essentielles. Elles offrent :

- Gestion des accès (Identity and Access Management, IAM). L'IAM permet de répondre à la question : qui peut faire quoi sur quelle ressource ? Garantissant ainsi que seuls les utilisateurs authentifiés et autorisés peuvent accéder aux données et aux ressources sensibles.
- Prévention de la perte de données (Data Loss Prevention, DLP). La DLP constitue un ensemble de règles métier qui classifient les données de l'entreprise afin de protéger les données sensibles contre l'exfiltration, les accès, les partages ou encore les modifications non autorisées.
- Sécurité du réseau cloud (Cloud Network Security). La sécurité du réseau cloud protège les réseaux basés sur le cloud des cyberattaques, des intrusions et des violations de données en filtrant le trafic, en appliquant des contrôles d'accès et en détectant les activités malveillantes.
- Gestion des informations et des événements de sécurité (Security Information and Event Management, SIEM). Le SIEM collecte et analyse les données de sécurité provenant de multiples sources, permettant aux organisations d'identifier et de répondre aux menaces rapidement.
- Renseignements (Threat Intelligence). Le renseignement fournit aux organisations des informations en temps réel sur les menaces émergentes, leur permettant de s'attaquer de manière proactive aux risques potentiels et de protéger leurs systèmes.

A cela, il faut aussi ajouter que les organisations ayant mis en place une pratique FinOps (idéalement, toute entreprise allant dans le Cloud, donc) proposent des configurations pré-approuvées (donc plus sûres) et surveillent l'usage des ressources en temps réel. **Les équipes FinOps et Sécurité ont donc tout intérêt à collaborer pour se nourrir l'une l'autre des bonnes pratiques**.

#Tech #Cloud

Cloud : y aura–t-il encore des “moments iPhone” ?

Prédire qu’il n’y aura plus de grande innovation dans le cloud serait exagéré, mais…

Au cours des dernières années, l'avènement du cloud public a été marqué par de nombreuses innovations qui ont radicalement transformé les usages et la façon de produire de la valeur pour les entreprises. On pense bien sûr aux offres PaaS, toujours plus étoffées, mais aussi à la démocratisation des offres serverless et la simplification de la mise en œuvre d’architectures découplées, mettant à profit les concepts modernes comme les microservices ou l’orientation événement.

Néanmoins, force est de constater que le rythme des annonces de ce type de produit a sensiblement baissé au cours des derniers mois.

Plusieurs raisons à cela :

- La marché arrive à un certaine forme de maturité, où les “briques de base” du cloud sont considérées comme des commodités. Les entreprises ont maintenant besoin de continuer à bâtir sur ce “new normal” ainsi que de finaliser leur modèle opérationnel qui a fortement chamboulé leurs habitudes.
- La vague GenAI a tout emporté sur son passage, et a certainement mobilisé une bonne partie des efforts des fournisseurs de cloud dans le domaine de la R&D - les annonces faites lors de leurs conférences respectives ne laissent guère de place aux doutes.

Mais cela ne signifie pas pour autant que le Cloud Public ne soit plus synonyme d'innovation, bien au contraire. L’exhaustivité, la maturité des services offerts, l’apport immense de l’IA générative dont on ne voit que les prémices aujourd'hui, font que le Cloud Public est, et restera, la plateforme de choix des entreprises pour accélérer leur développement et prendre un avantage concurrentiel.

#Tech #Cloud #clouddeconfiance

2024 : le vrai départ du Cloud de Confiance

L’alliance de l'hyperscaler et de l'hyperlocal

Les tensions géopolitiques créent une demande croissante pour des fournisseurs et des solutions cloud locaux. Les entreprises souhaitent avoir un plus grand contrôle sur l’accès à leurs données et leur infrastructure, et elles sont préoccupées par la sécurité de leurs données et de leurs traitements en cas de conflit géopolitique.

Le cloud de confiance est une solution qui répond à ces besoins. Il offre une combinaison de sécurité et de performance qui est adaptée aux entreprises qui souhaitent déplacer leurs données et leurs applications vers le cloud en garantissant un certain niveau de souveraineté.

Une entreprise préoccupée par ces aspects pour tout ou partie de ses données et traitements devrait considérer cette possibilité de recourir à un Cloud de Confiance, tel qu’il est défini par les autorités françaises.

Le cloud de confiance est un concept encore relativement nouveau, mais qui gagne rapidement en popularité - et dont les offres commencent véritablement à voir le jour. Il est soutenu par les gouvernements de nombreux pays, qui souhaitent promouvoir la souveraineté numérique de leurs entreprises.

Il existe deux types principaux d'acteurs cloud locaux briguant ce label :

- Les acteurs qui proposent principalement des datacenters avec des solutions qui sont très proches du hardware. Ils sont utiles pour les entreprises qui souhaitent déplacer leurs données et leurs applications vers le cloud, mais qui souhaitent également conserver un contrôle étroit sur leur infrastructure.
- Les acteurs locaux qui s'appuient sur des technologies d'hyperscaler (et nouent donc des accords avec Google, AWS, Azure). Ils proposent un panel de solutions plus large, qui est plus proche de ce que l'on attend d'un fournisseur cloud traditionnel. Ils sont utiles pour les entreprises qui souhaitent bénéficier des avantages du cloud, tels que l'agilité et l'évolutivité, tout en conservant un certain contrôle sur leurs données.

SecNumCloud de l'ANSSI permet un niveau de sécurité parmi les plus élevés du marché

#Tech #Cloud #FinOps #GreenOps

Equivalence FinOps - GreenOps : la fin du flou ?

Pendant longtemps, les observateurs et acteurs du cloud, dont nous faisons partie, ont considéré que FinOps et GreenOps étaient grosso modo équivalents, ou a minima les deux faces d'une même pièce. Et que la pratique FinOps amenait donc de manière systématique des gains "environnementaux".

D'une certaine manière, cela reste absolument vrai, par exemple dans les cas où l'on réarchitecture des applications en utilisants les offres telles que les Functions-as-a-Service ou les containers. En appliquant les mêmes principes que ceux régissant la virtualisation, à savoir un usage plus efficient des ressources, on diminue à la fois la facture et l'empreinte carbone. En tout cas il s'agit de quelque chose de communément admis. Il en va de même pour l'arrêt des services non utilisés.

Vers plus de transparence en 2024

En revanche, il manque des chiffres pour quantifier tout cela, faute d'informations précises de la part des cloud providers - c'est là que le bât blesse. Qui est en mesure de dire ce que font les cloud providers une fois les ressources utilisateurs éteintes ? Les serveurs physiques restent-ils allumés, sont-il optimisés pour accueillir d'autres charges de travail ? Cela reste du ressort du cloud provider et hors de notre périmètre d'utilisateur.

De la même manière, l'optimisation financière pure (réservation, exécution de charge de travail dans des régions moins chères) n'est pas systématiquement synonyme de réduction de l'empreinte carbone - par exemple si le nombre de machines utilisées reste sensiblement le même.

Que peut-on attendre de 2024 ? La généralisation d'outils de suivi différenciés entre FinOps et GreenOps, l'émergence de bonnes pratiques spécifiques aux deux pratiques, et un accroissement de la transparence des chiffres fournis par les cloud providers dans le domaine des émissions de gaz à effet de serre.

#Tech #Cloud #AIOps

AIOps : l’introduction de l’IA dans les opérations

Introduite pour la première fois par Gartner en 2016, la notion d''AIOps n’a pas vraiment percé depuis. Elle revient aujourd’hui, boostée par l’IA générative : ce serait formidable de pouvoir superviser son IT en dialoguant… En réalité, l’AIOps reste ancrée dans le big data et l'apprentissage automatique pour automatiser les processus opérationnels informatiques, notamment la corrélation des événements, la détection des anomalies et la détermination de la causalité. Ce qui change, c’est qu’on gagne en maturité sur le sujet et qu’on dispose de davantage de données.

Les pratiques AIOps sont particulièrement pertinentes pour les applications critiques où les attentes des clients sont très élevées, mais aussi quand il s’agit de traiter des grosses volumétries de données en temps réel ou quasi temps réel, qui proviennent de plusieurs sources avec des formats hétérogènes.

L’AIOps repose sur un certain nombre de piliers comme l’observabilité, la scalabilité, la fiabilité, l’automatisation, la précision ou encore l’explicabilité.

Les cas d’usage de l’AIOps

Concrètement, on peut dénombrer quelques cas d’usage de l’AIOps :

- **L’analyse des performances** pour optimiser l’utilisation des ressources et améliorer l’expérience utilisateur.
- **La détection proactive d’anomalies** et réduction des faux négatifs ou des faux positifs.
- **L’analyse et la corrélation des événements** afin de les contextualiser et de réduire le bruit.
- **L’automatisation de la gestion des services IT** pour réduire au maximum les tâches manuelle et répétitives et à faible valeur ajoutée pour se concentrer sur l’essentiel.

Afin de commencer à utiliser l’AIOps, deux options se présentent : la première peut être disponible très rapidement au travers des solutions sur étagère proposées par les éditeurs de logiciels ou les fournisseurs cloud, qui pour la plupart ont intégré des fonctionnalités AIOps dans leurs solutions. La seconde serait de construire sa propre solution en l’absence de logiciel du marché qui puisse répondre aux besoins spécifiques de l’entreprise. Par contre, il faudra bien réfléchir avant de se lancer dans un tel chantier qui peut s’avérer fastidieux et très compliqué à mettre en œuvre - sans parler de la maintenance et de l’évolution.

ORGA

/02

ÖRGA

LES CONTRIBUTEURS DE CETTE SECTION

ILS ET ELLES ONT ÉCRIT LES TECH TRENDS 2024

Sandrine BOITEAU
Consulting Director Orga & Product,
WENVISION

Aude DEFRETIERE
Consultante Product Strategy,
WENVISION

Romain MAZUIR
Head of Product Strategy,
WENVISION

Vanessa PERILLAT
Group Chief Talent Officer,
SFEIR | WENVISION

Olivier RAFAL
Consulting Director Strategy,
WENVISION

ORGA
#Equipes

De nouveaux modèles de travail en équipe

#Orga #Equipes #C4E

Généraliser les C4E pour porter votre dynamique

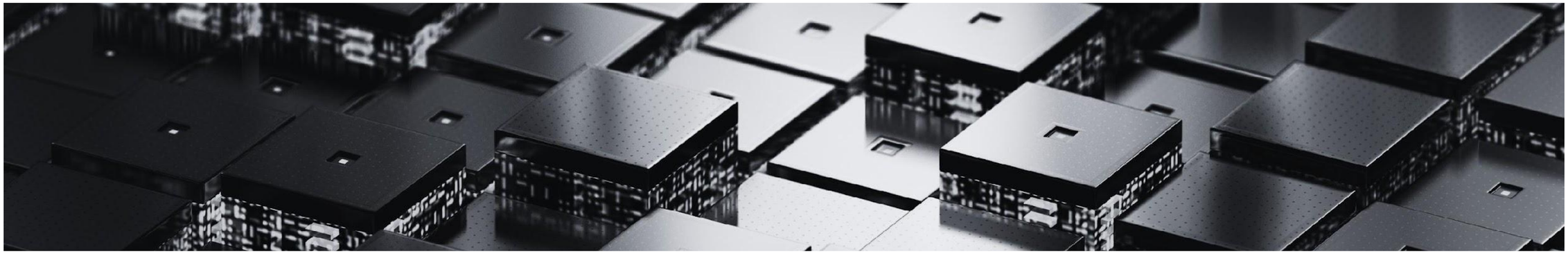
Les centres d'excellence (CoE, pour Centers of Excellence) classiques centralisent les experts et les assets d'une compétence donnée (architecture, sécurité, cloud, agilité, technologie...). Ils traitent généralement un domaine de problème spécifique avec des solutions novatrices et favorisent le succès d'un produit ou projet spécifique.

Ils permettent aux entreprises de gagner en expertise, mais doivent être avant tout considérés comme temporaires, car ils peuvent être difficilement scalables.

L'enjeu actuel consiste à passer d'une organisation CoE à une organisation C4E (Center for Enablement). Les C4E définissent alors des patterns réutilisables pour répondre aux besoins des équipes. Ils jouent le rôle d'un centre de mentoring avec enrichissement décentralisé. Pour la gestion d'un produit, c'est l'équipe produit qui est constituée de champions (en architecture, en sécurité, en cloud, en agilité...) et non plus d'experts choisis et détachés d'un CoE. **Les C4E, par la force de l'intelligence collective, accélèrent la mise à l'échelle.**

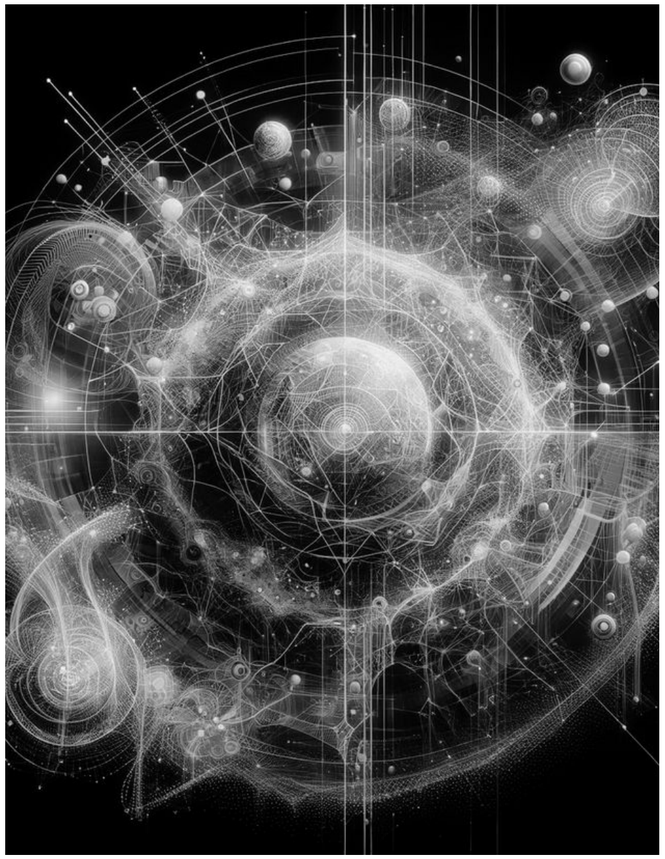
Problèmes très spécifiques et limités ou enjeux complexes ?

Le passage des CoE aux C4E devrait être une tendance plus importante dans le monde de la tech car cela favorise la dynamique par l'autonomie des équipes et l'adaptation des organisations aux changements rapides du marché.



#Orga #Equipes #complexité #cadre

Gérer la complexité entre global et ad-hoc par sous-système



Les futures entreprises efficientes sont celles qui donneront des cadres généraux, et pas *le* processus et *la* méthodologie communes.

Certes une organisation a besoin de pouvoir rester cohérente et synchronisée autour de la stratégie globale. Et traditionnellement, la définition de processus, méthodologies, procédures et outillages communs permet de garantir l'ordre. Mais quel est son niveau de compatibilité dans un monde qui change plus vite que le taux d'adaptation des processus ?

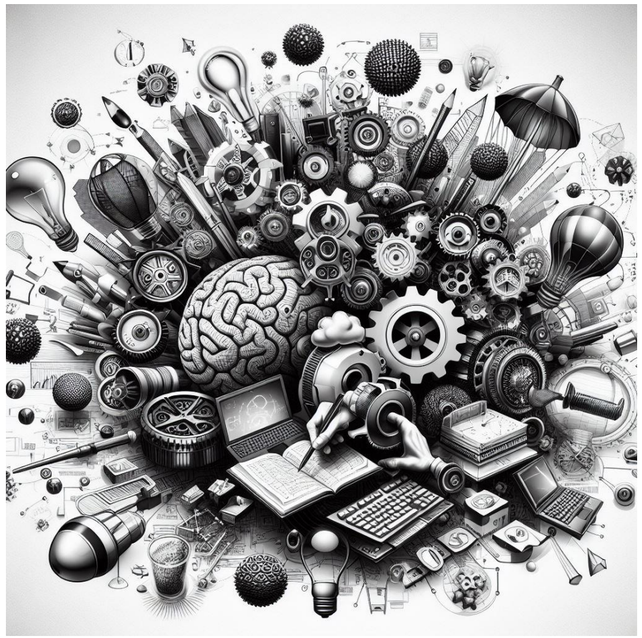
S'inspirer du corps humain

Dès lors, les entreprises ont besoin de gérer leur complexité, où chaque organe (projet, département, équipe, produit…) du corps global (l'organisation) sait s'auto-gérer, tout en étant connecté aux autres organes par des interactions transverses (réunions, pratiques, outils…). Il s'agit de formuler des principes globaux en transverse (« favoriser l'intelligence collective », « sélectionner une plateforme technologique permettant de communiquer avec des API »…) plutôt que des processus et solutions prédéfinis (« faire 1 réunion par semaine », « Java comme seul langage pour toute l'entreprise »…). Les patterns vont ainsi répondre à ces besoins globaux de cohérence, d'alignement avec la stratégie (OKR de plus haut niveau) et favoriser la sélection et l'implémentation de solutions ad hoc, permettant aux sous-écosystèmes de l'entreprise de s'auto-organiser.

Afin de conserver une cohérence, la généralisation des Centres for Enablement (C4E) ira de pair avec l'usage de patterns. Le pattern et son implémentation peut alors suivre une courbe d'adoption, facilitée par les C4E : découverte, qualification, definition, pilote, standard, industrialisation, divergence, décroissance, suppression.

#Orga #Equipes #dataproduct #engineering

Data Product Engineering : bien intégrer la data avec avec le SI



Comment délivrer des produits Data intégrés? Nous pensons que le concept de Data Product Engineering doit émerger : une organisation Data qui utilise les les mêmes leviers d'ingénierie qu'une organisation Produit.

Il n'y a aucun doute sur le fait que la data soit devenue un actif stratégique pour les entreprises. Elle est utilisée pour prendre des décisions, améliorer les produits et services, et créer de nouveaux modèles économiques.

Ce qui est moins évident, c'est que la data est un produit comme un autre. Elle doit être collectée, traitée et valorisée de manière efficace et efficiente. Et elle dépend aussi de son stockage, des capacités du réseau, de l'architecture, de normes réglementaires et de sécurité. Comme tout produit, sa production peut être industrialisée, au sein d'une Factory.

La complexité de ce produit, c'est qu'il nécessite une étroite collaboration entre des équipes data (qui connaissent la donnée et ses traitements), des responsables produits (finalité des produits data, cycle de vie, pilotage par la valeur…) et des équipes IT (plateforme et pipelines de données, outils de consultation…).

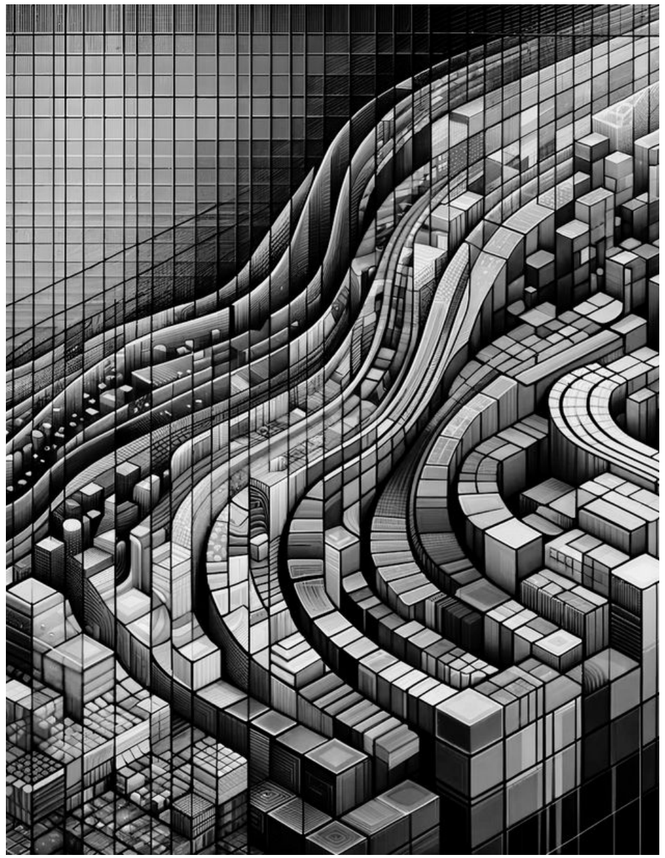
Chaque entreprise doit mettre en place un TOM adapté à ses caractéristiques

Comment assurer une collaboration étroite entre ces équipes et profils ? Malheureusement, il n'existe pas de méthodologie unique pour la data et la fluidification depuis le besoin jusqu'à l'usage en production est essentielle pour la réussite des produits data. Il revient à chaque entreprise de construire son modèle opérationnel (TOM, target operating model), adapté à son organisation et sa culture. Il existe tout de même des patterns, qui pourront être employés pour accélérer la configuration de ce TOM et son déploiement, et s'assurer de l'application de bonnes pratiques.

Dans tous les cas, nous constatons qu'après les initiatives de DataLab isolés, les Data Factories font face à leur dépendance avec le reste du SI - et c'est un signe de maturité. Dans ce contexte, les entreprises vont devoir **adopter une approche d'ingénierie de la data, s'appuyant sur des processus et outils de livraison bout en bout, ainsi que des principes inhérents aux organisations orientées produits, comme les approches Lean et Agile.**

#Orga #Equipes #pratiques #rôles

Le prêt-à-porter des pratiques, pas de taille unique



Peut-on encore, en 2024, poser les bases d'une organisation agile à l'échelle, avec un framework, des processus, une méthodologie commune à toute l'entreprise ?

La fin du principe de *La même pratique pour tous / One size fits all*

Nous avons plutôt tendance à guider nos clients vers des approches plus flexibles et adaptatives. C'est un critère essentiel pour la réussite des évolutions technologiques à adopter. Il faut l'admettre : les pratiques uniformes ne seront *plus* efficaces. La recherche de *la* méthodologie unique ou du *même processus pour tous*, pour toute l'entreprise, ne permet plus d'atteindre ni l'efficience requise, ni la satisfaction des clients et l'épanouissement des collaborateurs. La diversité des contextes requiert des approches flexibles et adaptables au sein même des organisations.

Les transformations d'organisation vers le Framework SAFe démontrent bien ce besoin de personnalisation, puisqu'on constate qu'il est rarement appliqué à la lettre. Non seulement les organisations adaptent SAFe en fonction de leurs besoins spécifiques, mais au sein de la même entreprise, chaque contexte (département, programme, projet, produit, équipe...) va également le modifier à ses besoins, engendrant des divergences de son application malgré les apparences.

Les entreprises efficientes d'aujourd'hui et de demain sont celles qui donneront des cadres, des patterns généraux, et non un processus et une méthodologie commune. Penser globalement et diverger localement.

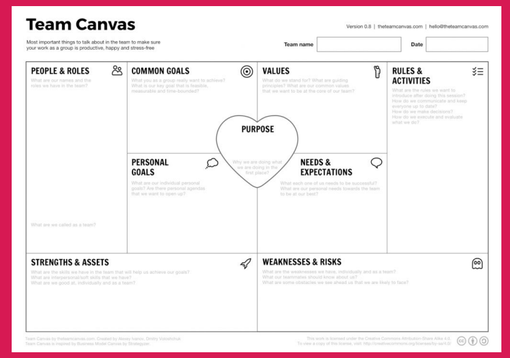
Vers une flexibilité des Rôles et Responsabilités

De même, avoir exactement la même définition des rôles dans toute l'organisation (Product Owner, Chef de Projet, PMO, Développeur, Ingénieur QA, Scrum Master, Agile Delivery Manager...) ne permet plus de s'adapter aux changements rapides du paysage économique et technologique en 2024. Pour cela, il devient crucial d'adopter une approche flexible permettant de faire évoluer ces rôles en fonction des besoins spécifiques de chaque équipe et écosystème, à des fins d'efficience dans un marché changeant rapidement, mais aussi pour satisfaire les collaborateurs qui vont bénéficier de plus de compétences et plus de possibilités - tout en conservant une structuration globale cohérente.

#Orga #Equipes #RACI

Quels rôles pour votre écosystème ?

Le Team Canvas : un espace visuel et collaboratif



Le modèle traditionnel de définition solide des rôles et responsabilités, soutenu par l'usage courant des matrices RACI (Responsibility, Accountability, Consulted, Informed), n'est plus adapté à la résolution des problèmes complexes auxquels les entreprises font face en 2024.

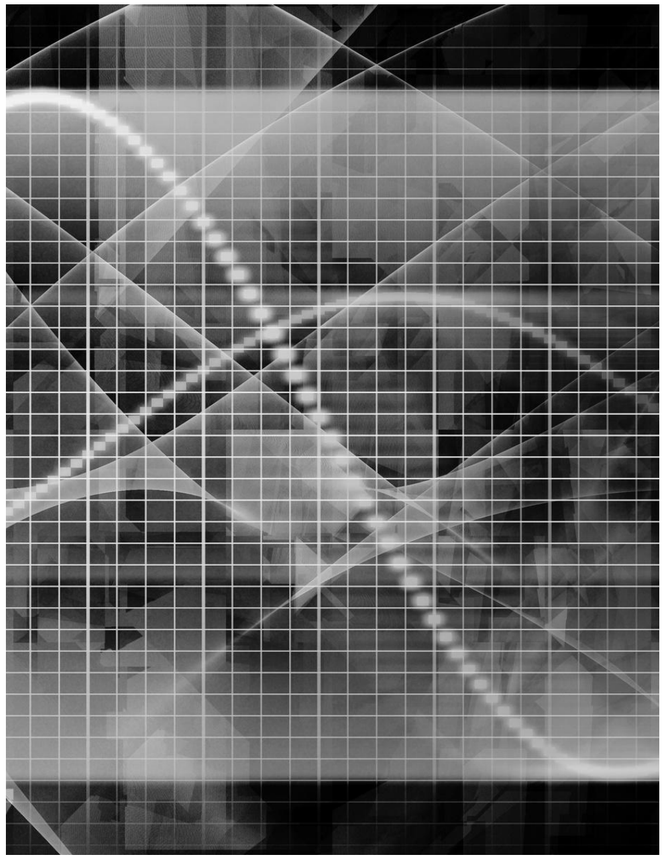
La résolution de problèmes complexes nécessite une approche collaborative, fondée sur le partage des responsabilités et le développement de compétences transversales et multiples dites T-Shaped. Ces pratiques favorisent l'épanouissement personnel et collectif, et permettent à chaque membre de l'équipe de contribuer de manière significative à la résolution des problèmes et à l'innovation.

Les rôles figés peuvent conduire à une perte de motivation et d'épanouissement des collaborateurs, alors que ce sont les responsabilités partagées qui contribuent à un environnement propice à l'épanouissement professionnel et personnel. Un espace visuel et collaboratif (comme le Team Canvas) facilite la responsabilité partagée et encourage la réflexion collective, l'itération constante. Ces adaptations en continu favorisent également la dynamique lean du flux de valeur.

Pour un écosystème performant, il faut **savoir ajuster en continu les rôles et responsabilités au sein d'une équipe** en fonction des besoins et tendances émergentes.

#Orga #Equipes #Agilite

De l’agilité à des approches lean personnalisées



L’agilité est devenue une évidence pour les organisations de toutes tailles et de tous secteurs. Cependant, elle est souvent vue comme un ensemble d'outils et de processus à respecter, alors qu'il s'agit avant tout d'un état d'esprit.

La mise en œuvre d'une méthode ou d'un framework agile est une étape importante, mais elle ne doit pas être considérée comme un objectif en soi. La finalité est de créer de la valeur pour le client et l'entreprise, en termes d'expérience utilisateur, de capacités d'innovation, de résilience et d'adaptabilité.

Par exemple, une entreprise qui adopte une approche agile peut développer des produits et services qui répondent mieux aux besoins de ses clients, en s'adaptant plus rapidement aux changements du marché.

L’agilité à l’échelle de l’entreprise : un état d’esprit, une culture commune au sein de toute l’entreprise

Pour ce faire, il est important de travailler sur le **découplage des équipes, en créant des équipes responsables de bout en bout d'un domaine**. Le succès de Team Topologies (livre, cours, conférences…) et de son utilisation avec le DDD (Domain Driven Design) montre des solutions pour des équipes qui n'ont pas à se synchroniser avec *n* équipes pour délivrer de la valeur.

Nous nous attendons à ce que les entreprises définissent leur **propre approche Lean, en puisant dans diverses approches qui ont fait leurs preuves**, comme de revisiter Kanban pour exploiter sa puissance.

De même, l'agilité doit concerner l'ensemble de l'entreprise et de ses services, pour atteindre les objectifs stratégiques (OKR, Objectives & Key Results). Les métiers doivent y être acculturés, mais aussi les dirigeants, ne serait-ce que pour piloter leur portefeuille de projets dans une perspective lean (Lean Portfolio Management).

L'agilité est un état d'esprit qui permet aux organisations de créer de la valeur de manière plus efficace. Elle doit être adoptée par l'ensemble de l'entreprise, et non pas seulement par le domaine IT.

#Orga #Equipes #OKR

Vers l’émergence de véritables OKR

En tant qu’outils pour aligner et coordonner les efforts d’une équipe, les OKR devraient obligatoirement figurer au menu des améliorations organisationnelles à apporter cette année pour gagner en efficience.

Introduite par Andy Grove chez Intel dans les années 1970 et popularisée par Google, la méthode OKR (Objectives & Key Results) est une stratégie de gestion visant à aligner les objectifs stratégiques avec des résultats opérationnels tangibles et mesurables. Elle est conçue pour encourager des ambitions élevées tout en nécessitant une mise en œuvre attentive.

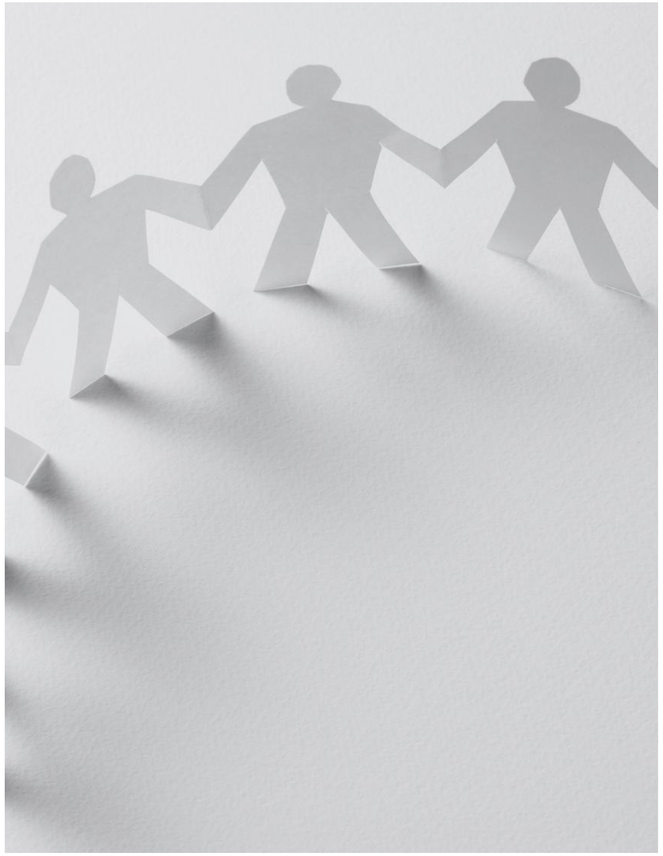
La définition des OKR n'est pas triviale. Ils doivent être en accord avec la mission et la vision de l'entreprise, sans devenir une liste exhaustive. L'importance réside dans la priorisation pour assurer clarté et efficacité. **Les OKR doivent être ambitieux pour motiver les équipes, mais aussi réalistes pour éviter la démoralisation**.

La méthode doit s'adapter à la culture et au système managérial spécifiques de chaque entreprise. Des ajustements pratiques sont donc souvent nécessaires pour l'alignement organisationnel. Tous les objectifs ne sont pas forcément quantifiables, certains peuvent être qualitatifs ou répondre à une simple logique binaire.

Les OKR sont une méthode de suivi des objectifs, et non une solution autonome de pilotage

La mise en place des OKR exige une maturité organisationnelle, impliquant un engagement envers l'alignement, la communication, et la gestion. Un système de gouvernance et un accompagnement initial des équipes sont essentiels pour une intégration réussie.

Les OKR ne remplacent pas nécessairement d'autres systèmes de mesure de performance, mais peuvent les compléter. Il est crucial de maintenir un équilibre avec les activités courantes de l'entreprise et de ne pas considérer les OKR comme une solution autonome. Leur succès dépend d'une gouvernance et d'une intégration organisationnelle adaptées. Tout en étant agile, il est préférable de conserver des objectifs stratégiques annuels pour une organisation efficace des équipes. Finalement, il est important de reconnaître que les OKR ne sont pas l'unique méthode de mesure de performance, et peuvent être complétés par d'autres outils tels que les KPI (Key Performance Indicators).



#Orga #Equipes #Recrutement

La to-do-list des équipes recrutement pour 2024

Pour être une tech company en mouvement, il faut continuellement upskiller son équipe recrutement. Voici les 3 grandes tendances pour 2024.

1- L'automatisation, prochaine étape du métier de recruteur

Il vaut mieux utiliser les recruteurs pour leur matière grise plutôt que pour faire des tâches répétitives et sans valeur ajoutée. Beaucoup d'outils sont déjà disponibles à cette fin sur le marché. L'avènement de l'IA générative - avec la course à la productivité qu'elle suscite - met en exergue ces solutions. Cela va sensibiliser les comités de direction et permettre ainsi de débloquer du budget. La plateforme Linkedin Recruiter va par exemple intégrer un ChatGPT spécial sourcing afin de permettre aux recruteurs d'aller plus vite dans l'**identification de profils pertinents** qui pourraient avoir un comportement sur la plateforme qui laisse entendre qu'ils sont à l'écoute du marché. L'automatisation va aussi être clé concernant la **gestion des tâches administratives**, comme les prises de rendez-vous, le suivi de candidatures, la recherche d'anciens candidats, l'écriture des comptes-rendus, l'écriture et la signature des contrats, etc.

2- Focus sur le marketing

La notion de "Talent Attraction" devient essentielle. Comment attirer des candidats ? Comment les atteindre avec des messages plus pertinents ? Quelle est la Value Proposition de l'entreprise & d'un job en particulier ? Comment pitcher efficacement l'entreprise - le storytelling - en s'appuyant sur sa culture d'entreprise ? Quelle stratégie marketing mettre en place afin de toucher les candidats passifs ? Là aussi, l'IA Generative et l'automatisation vont être de mise. Des solutions SaaS de plateformes candidats permettent d'automatiser de bout en bout les échanges, en étant multicanales (inmail, email, sms, messagerie linkedin, etc.) et en semi-automatisant les envois de messages d'approche et les relances. On commence à voir des assistants GenAI afin d'aider les recruteurs à écrire **des messages personnalisés en prenant en compte de nombreuses variables**.

3- Rechercher les soft skills adéquates pour évoluer plutôt qu'un large panel de hard skills

La Tech va tellement vite que, selon le rapport Future of Jobs du World Economic Forum, pour les personnes qui évolueront sur le même poste, 40% de leurs compétences de base changeront d'ici à 2025. **S'assurer que son entreprise disposera des bons profils dans les années à venir nécessite donc un changement de mindset : recruter pour des compétences et des capacités, plutôt que sur les performances passées**. Il va donc s'agir de prioriser les compétences en faisant le distinguo entre celles indispensables et celles nice to have. On parle du **concept de "transferable skills" : rechercher des personnes à potentiel**, qui n'auront pas toutes les compétences dès le début mais qui auront la capacité de les développer rapidement.

ORGA
#CareerPaths

Des rôles et chemins de carrière qui s'affinent

#Orga #CareerPaths

De nouveaux chemins de carrière pour les développeurs

La discipline du développement informatique est relativement récente. Il était donc naturel que les développeurs montent en grade, encadrent des projets, voire gèrent des équipes. Et ce faisant, développent de moins en moins. Si ce schéma convient très bien à certaines personnes, d'autres peuvent se sentir très frustrées de ne pouvoir évoluer qu'en s'éloignant du code et de la technique. Enfin certaines personnes n'ont tout simplement pas les compétences et/ou l'envie de gérer des projets ou des collaborateurs.

Nous anticipons pour cette raison une montée en puissance des contributeurs individuels, des experts dans leur domaine, reconnus comme tels.

Une branche pour des profils plus orientés management, une branche pour la technique pure

Cela implique la création de chemins de carrière spécifiques, avec une branche pour des profils hybrides techniques et managériaux, et une branche pour des profils purement techniques, qui pourront s'épanouir dans ce rôle. Dans chaque branche, les collaborateurs doivent disposer de ressources (temps, budget voire primes) pour mener à bien leurs missions d'encadrement, de veille, de prise de parole, etc. Ces chemins de carrière comprennent aussi des "redevabilités" : un ensemble de choses attendues de la part des collaborateurs, pour le bien de leurs équipes et de l'entreprise.

Car une entreprise qui accepte de dégager du temps à ses collaborateurs pour de tels engagements investit pour l'avenir :

- elle gagne en visibilité sur le marché : alors qu'il est difficile de recruter de bons profils, ces derniers seront davantage attirés par des pairs s'exprimant en public, témoignant du caractère pointu, innovant, des produits développés en interne et des technologies mises en œuvre ;
- elle offre des perspectives à ses populations techniques, qui peuvent se projeter et envisager d'y poursuivre une bonne partie de leur carrière ;
- elle propose un cadre cohérent dans lequel l'ensemble des collaborateurs techniques pourra se retrouver, même dans un contexte international et décentralisé.

#Orga #CareerPaths #RH

Des modèles T-Shaped pour s’adapter aux nouveaux défis Tech

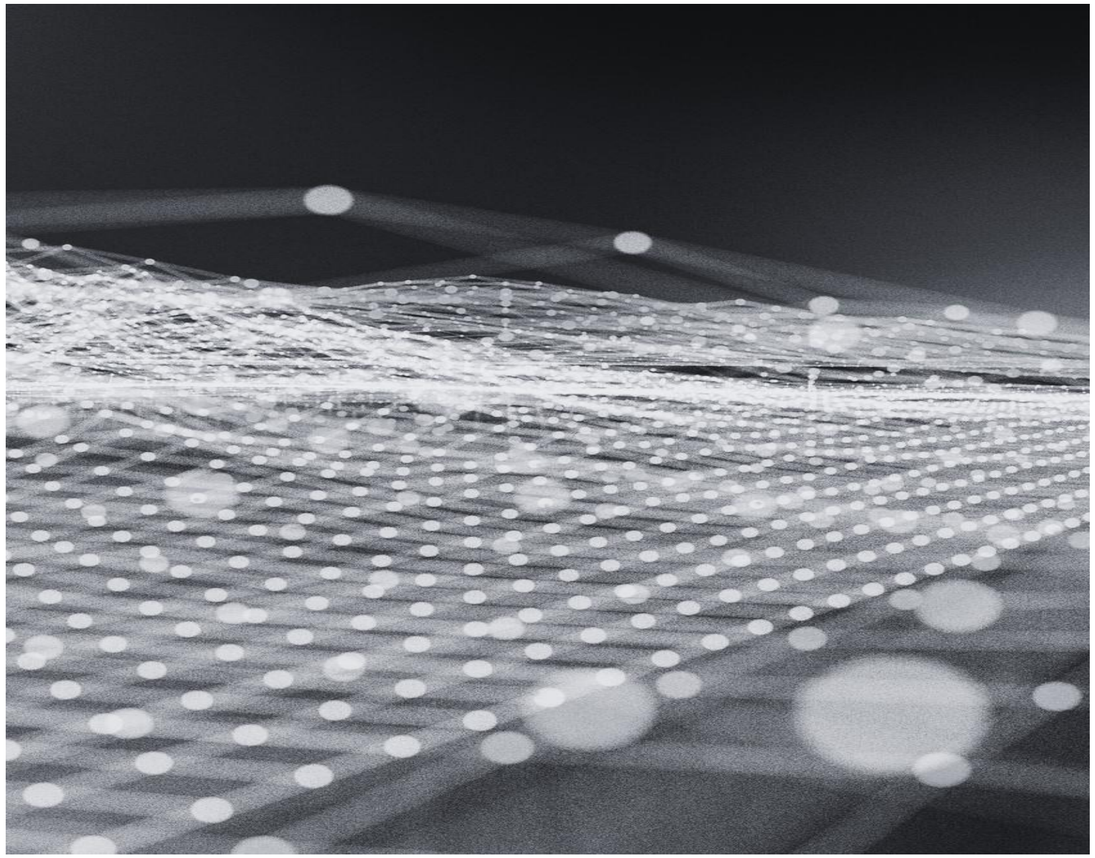
Les nouveaux modes de travail exigent des compétences variées et une adaptabilité constante. Comment répondre aux besoins de talents en Cloud Computing, en Intelligence Artificielle, en Machine Learning, ou toute autre prochaine tendance ?

Les fiches de poste traditionnelles ne reflètent plus la réalité des rôles évolutifs. Par exemple, un développeur mobile peut avoir besoin de développer son expertise fullstack pour répondre aux besoins d’une squad. De la même manière, un data scientist peut avoir besoin de développer ses compétences en Data Engineering Machine Learning pour travailler sur des projets d'Intelligence Artificielle.

Aider les profils tech à anticiper les évolutions du marché et se former en conséquence

Le modèle T-shaped selon lequel les individus ont une expertise approfondie dans une discipline, tout en étant compétents dans d'autres domaines, favorise la flexibilité et la polyvalence. Mais ce modèle est encore méconnu des RH. Les responsables des ressources humaines doivent s’adapter pour faire face aux besoins de changement permanent, en s'appropriant un état d'esprit agile et collaborer étroitement avec les responsables et les managers. Ils doivent également comprendre les nuances des nouveaux rôles, identifier les compétences essentielles, créer et tester des programmes de formation ad-hoc alignés sur des besoins évolutifs.

Le rôle actif des RH sera essentiel dans l'accompagnement des compétences des équipes en 2024. Les RH devront jouer un rôle de facilitateur, en aidant les employés à développer leurs compétences et à s'adapter aux nouveaux modes de travail et aux nouvelles tendances technologiques.



#Orga #CareerPaths #CTO

Le CTO est l’interlocuteur privilégié du CEO pour les aspects IT

Le rôle de CTO est relativement nouveau dans les entreprises. En l’absence de définition universelle de son poste, et vu que la technologie évolue constamment, l'étendue des fonctions du ou de la Chief Technology Office évolue aussi en permanence.

Le rôle de CTO pour prendre acte du statut de tech company

On peut dégager deux éléments qui doivent caractériser ce rôle de CTO :

- Le ou la CTO fait partie de l'organe de gouvernance de l'entreprise (ComEx, Comité exécutif...). C'est un C-Level.
- Il/elle possède une forte expertise technologique dans plusieurs domaines.

Dans une start-up, le CTO est un allié du CEO, qui va créer toute la partie technique et faire grandir ses équipes. Lorsque l’entreprise grossit, le CTO peut prendre du recul, donner les directives, résoudre les problèmes épineux. Et toujours discuter business avec le CEO.

Un CTO Office afin de pouvoir piloter tous les sujets stratégiques

La croissance de l’entreprise verra arriver un Chief Product Officer, pour guider aux destinées des produits. Le CTO prendra le plus souvent la direction technique de la plateforme, ou bien la laissera aux bons soins de son CTO Office ; lui sera alors l’égérie technique de l’entreprise - et toujours l’interlocuteur privilégié du CEO sur la manière de bénéficier des capacités IT pour le business.

Le rôle de CIO pour les outils du SI interne

Alors que le CTO travaille sur les assets technologiques permettant d’offrir des produits de qualité aux clients, le CIO s’assure que les collaborateurs peuvent travailler efficacement. Il planifie, définit, sécurise et supervise le SI interne, qui permettent aux processus de l’entreprise de fonctionner correctement.

Traditionnellement, le CIO gère les fournisseurs et les achats, les équipes d’intégration et de support, ainsi que la réalisation de certaines applications spécifiques. Dans une entreprise moderne, le CIO sera en quelque sorte le CTO du SI, la personne qui construit à destination de ses clients internes un SI moderne, compatible avec le futur.

Le rôle de VP Engineering pour gérer les talents et les équipes

La personne qui gère la réalisation d’un produit ou d’une plateforme ne devrait pas être le manager hiérarchique des équipes de développement : les intérêts d’un patron de produit, CPO ou CTO, et ceux d’un développeur peuvent considérablement diverger. La carrière du développeur sera davantage considérée par un manager détaché de l’opérationnel.

C’est pourquoi la fonction de VP Engineering prend corps aujourd’hui : il s’agit de recruter, accompagner, former des équipes pluridisciplinaires, faire en sorte qu’elles grandissent et s’épanouissent, se forment sur d'autres technologies, et aient envie de mettre à profit leurs nouvelles compétences au service de l’entreprise.

CULTURE

/03

CULTURE

LES CONTRIBUTEURS DE CETTE SECTION

ILS ET ELLES ONT ÉCRIT LES TECH TRENDS 2024

Sandrine BOITEAU
Consulting Director Orga & Product,
WENVISION

Mylène MAIGNANT
Consultante IA,
WENVISION

Feliks PLAWCZYK
Architecte, Consultant,
WENVISION

Olivier RAFAL
Consulting Director Strategy,
WENVISION

CULTURE #GrandsPrincipes

Davantage d’autonomie, via l’acculturation et l’agilité

#Culture #GrandsPrincipes #agilité #devsecops

Une agilité holistique pour des organisations résilientes

L'agilité est une nécessité pour les entreprises qui souhaitent réussir dans un monde en constante évolution. Elle permet aux entreprises de s'adapter rapidement aux changements, de répondre aux besoins des clients et de créer de la valeur de manière durable.

En 2024, l'agilité continuera d'être une tendance majeure dans le monde de l'entreprise. Avec quelques nuances essentielles qui s'observeront chez les entreprises les plus en pointe :

- **L'agilité s'étendra à l'ensemble de l'entreprise**. L'agilité ne se limite plus au domaine IT. Elle est désormais adoptée par les équipes de vente, de marketing, de finance, des ressources humaines...
- **L'agilité sera plus holistique.** Les entreprises intégreront les principes agiles à leurs processus, à leur culture et à leur stratégie.
- **L'agilité sera plus data-driven.** Les entreprises utiliseront des données pour mesurer les résultats de leurs initiatives agiles et pour améliorer en continu leurs processus.
- **L'agilité s'appuiera sur l'automatisation et l'IA.** Les technologies peuvent aider les équipes à automatiser les tâches répétitives, à prendre des décisions plus rapidement et à améliorer la qualité des produits et services.
- **L'agilité servira la cybersécurité.** L'agilité permet aux équipes de réagir aux menaces émergentes et de mettre en œuvre des correctifs de manière efficace autour de la culture DevSecOps.
- **L'agilité s'appuiera sur le cloud.** La montée en puissance du cloud permet aux équipes de travailler de manière plus flexible et collaborative. Le cloud offre une plateforme scalable et accessible de n'importe où, ce qui est essentiel pour les équipes agiles qui travaillent souvent sur des projets à l'échelle mondiale.

#Culture #GrandsPrincipes #Acculturation

Acculturer est la clé du succès de l’IA générative en entreprise

“Qu’est-ce que l’IA ? Le Machine Learning et l’Intelligence Artificielle sont-ils similaires ? Quel est le futur de l’IA, va-t-elle remplacer les humains ?” Avant de travailler sur l’implémentation d’une IA générative au sein d’une entreprise, il est essentiel d’aligner l’ensemble des participants au projet sur ces questions.

Déconstruire le mythe de l’IA

Si l’IA est omniprésente dans les médias aujourd’hui, de nombreuses interrogations quant à son fonctionnement restent pourtant en suspens et compliquent son adoption. Les journaux mais aussi les films et les romans tendent à véhiculer une vision fantasmée des modèles algorithmiques. Ce qui contribue à provoquer une certaine réticence de ceux à qui on demande de l’intégrer dans leurs pratiques professionnelles quotidiennes.

Exposer son fonctionnement de manière pédagogique

La première étape consiste à proposer des explications claires et pédagogiques à l’ensemble des collaborateurs. Grâce à des formations adaptées au niveau du public, il s’agit d’éclaircir les mécanismes de la GenAI. Clarifier ce que signifie un algorithme, expliquer les similitudes mais aussi les différences entre la machine et l’humain, répondre à toutes les questions éthiques, techniques ou juridiques sur l’IA : autant de thématiques à aborder dans un cadre bienveillant pour désacraliser cette technologie et préparer le terrain pour une adoption sereine et profitable.

Méconnaissance et rétiences se rencontrernt aussi bien chez les utilisateurs métiers, pas nécessairement familiarisés avec les sujets IT, que chez des développeurs aguerris, peu enclins à accepter qu’une IA puisse réaliser une partie de leurs tâches. C’est pourquoi il est crucial de souligner tant les possibilités que les limites de l’IA durant ces formations. L’IA n’est pas une baguette magique. Elle ne peut remplacer l’humain, elle en a d’ailleurs besoin pour fonctionner correctement. Nuancer les capacités de cet outil permettra à la fois de désacraliser sa puissance et d’apaiser certains esprits inquiets face à son essor.

Proposer des ateliers pratiques

La seconde étape d’une acculturation réussie repose sur des exercices pratiques qui permettront aux participants de mieux saisir les bénéfices de l'IA générative. Si on peut lire qu’elles aident à la rédaction de mails, construisent des supports de présentation visuelle et synthétisent des documents toutes seules, encore faut-il maîtriser l’art du prompt pour obtenir des résultats pertinents et convaincants. De la même manière, si l’IA générative est performante pour certains types de tâches, elle le sera moins pour d’autres. Il est donc primordial d’identifier les cas d’usage dans lesquels une telle technologie peut apporter une plus-value.

#Culture #GrandsPrincipes #Liberté #Responsabilité

Vers l’autonomisation des collaborateurs

Pourquoi fait-on les choses ? C’est une question qu’on se pose encore malheureusement trop souvent dans les entreprises. Collectivement, parfois, lorsqu’on lance des projets IT dont on ne sait pas vraiment quelle valeur business cela pourra créer. Individuellement, souvent, lorsqu’un collaborateur doit accomplir une tâche, travailler sur un projet, sans connaître ou sans véritablement en comprendre la finalité.

La capacité des entreprises à innover, à être performantes, dépend pourtant directement de la performance individuelle des collaborateurs, de leurs compétences et de leur capacité à prendre des initiatives qui contribueront à la stratégie globale. Cette vision stratégique doit donc être définie, partagée, avec des collaborateurs auxquels il faut faire confiance pour exécuter cette stratégie.

C’est la clé de la motivation que de donner du sens : les frameworks d’agilité, les enseignements de John Kotter ou dernièrement le livre No Rules Rules d’Erin Meyer et Reed Hastings (le PDG de Netflix) insistent tous sur cette nécessité ; prenons le pari qu’au vu des remous économiques que nous devond traverser, les entreprises auront à cœur de donner plus d’autonomie à leurs collaborateurs, pour leur propre bien.

Le corollaire sera bien sûr la nécessité pour les collaborateurs de faire de la veille quotidiennement, de se former, de manière proactive, sur les vagues technologiques successives. Il n’y a pas que les entreprises et leurs systèmes d’information qui doivent constamment s’adapter : les collaborateurs doivent adopter une culture de l’apprentissage en continu.

CULTURE
#Pilotage&Ethique

Quel état d’esprit adopter face aux grands challenges de 2024

#Culture #Pilotage&Ethique #ValueDrivenIT

Le pilotage par la valeur, pour sortir de l'IT centre de coûts

Le numérique est un formidable vecteur de création de valeur. Mais le secteur informatique dans son ensemble a toujours été très mauvais pour le faire comprendre. L'avènement des Tech Companies (dans la finance, les transports, la santé…) a montré à tous, dirigeants et salariés, la puissance de l'IT quand elle génère directement du business.

Le problème, c'est que l'ensemble des investissements nécessaires pour obtenir des résultats visibles paraît lui très abstrait, intangible. Typiquement, comment faire comprendre à un responsable métier, un dirigeant, qu'il est nécessaire d'investir dans des compétences, dans une infrastructure cloud, un framework, pour délivrer une application d'IA générative d'entreprise, pertinente et sécurisée, lorsqu'on a l'impression de pouvoir obtenir la même chose gratuitement en ligne ? Comment expliquer que le développement initial ne représnte qu'une partie du coût ?

C'est tout un changement organisationnel et culturel qu'il convient d'initier. Chacun dans l'entreprise, y compris au sein de l'IT, doit embrasser cette vision d'une IT au service de la création de valeur. Il s'agit de sortir d'un mode projet traditionnel, qui consiste à demander des moyens (le fameux IT centre de coûts), au profit d'un mode produit, où on investit pour créer de la valeur.

Passer à l'orientation produits et au FinOps

Cela demande donc de revoir l'organisation, pour passer en mode produit et discuter des investissements avec les product owners. Cela demande aussi des pratiques et de l'outillage adapté, relevant du FinOps, une discipline qui vise justement à rendre les coûts transparents et le plus prévisibles possible, de façon à pouvoir ajuster les investissements nécessaires en fonction de la valeur business qu'on cherche à créer, exprimée sous forme de KPI (chiffre d'affaires, nombre d'utilisateurs, rapidité d'exécution…).

#Culture #Pilotage&Ethique #Souverainete

La souveraineté, un horizon qui le restera

La notion de souveraineté revient depuis des années dès lors qu'on aborde le sujet du Cloud. C'est une question tout à fait légitime. Quid en effet de la capacité d'un État, d'une entreprise à pouvoir bénéficier d'un service, d'une technologie, de ses données, quels que soient les aléas externes ? C'est pourquoi d'ailleurs nombre d'entreprises continuent de vouloir gérer elles-mêmes leurs serveurs, dans leurs datacenters. Leurs données semblent ainsi à l'abri. Mais peut-on se considérer souverain quand on assemble un serveur d'un fabricant américain tournant sous un OS américain avec des composants et des matériaux issus d'un peu partout dans le monde, un routeur chinois sous un OS chinois, des logiciels propriétaires dont l'usage n'est garanti que tant qu'on paie la maintenance (l'éditeur pouvant couper l'accès à tout moment), etc. ? L'Open Source constitue un des éléments de réponse, mais il ne peut pas tout, et il reste tout l'environnement matériel… Car la souveraineté ne s'applique pas qu'aux données, elle s'étend aux traitements.

C'est pourquoi il vaut peut-être mieux considérer la souveraineté comme un horizon qu'on cherche à atteindre, et faire dès aujourd'hui un certain nombre de choix pragmatiques. En commençant par considérer que la souveraineté numérique est un sujet à traiter au niveau des Etats, comme c'est le cas aujourd'hui des questions militaires, énergétiques ou alimentaires.

Des solutions spécifiques pour les OIV et les OSE

Hors obligations spécifiques (pour les OIV et OSE, opérateurs d'importance vitale et opérateurs de services essentiels) et contraintes règlementaires (RGPD, ACPR, etc.), les entreprises doivent de toute façon garantir à leurs clients, partenaires ou actionnaires qu'elles emploient tous les moyens à leur disposition pour être les plus efficientes possible tout en assurant la sécurité et la confidentialité de leurs données et des traitements associés.

Dans la grande majorité des cas, cela revient à utiliser le cloud public pour bénéficier de ses apports (performance, sécurité, efficience énergétique…). Quant aux exigences légales qui s'appliquent aux OSE et OIV, les solutions "cloud de confiance" apportent une réponse adéquate pour des données et traitements spécifiques (à ne pas confondre avec les solutions dites souveraines, qui ne reposent sur aucun fondement juridique).

Prendre ce type de décision demande un pilotage fin, alliant connaissance des contraintes juridiques, compréhension des enjeux technologiques et des offres du marché, et beaucoup de doigté pour faire avancer un sujet souvent éminemment politique.

#Culture #Pilotage&Ethique #Sustainability

L’écoconception et la Tech for good s’ancrent dans le paysage

Il n’est pas nécessaire de se flageller à coups de chiffres sortis du chapeau sur le poids d’un email, d’un tweet ou d’une vidéo pour prendre conscience que l’IT consomme énormément de ressources : électricité bien sûr, mais aussi des composants dont il faut extraire les matières premières, du mazout pour les faire voyager par mer, etc. La question n’est donc pas tant de savoir si les externalités positives (réduction des déplacements, création de richesse, accès à la culture…) compensent les négatives, mais plutôt de faire en sorte que notre consommation de l’IT se fasse de façon raisonnable et raisonnée, pour créer de la valeur, sur le plan économique bien sûr, mais aussi sociétal.

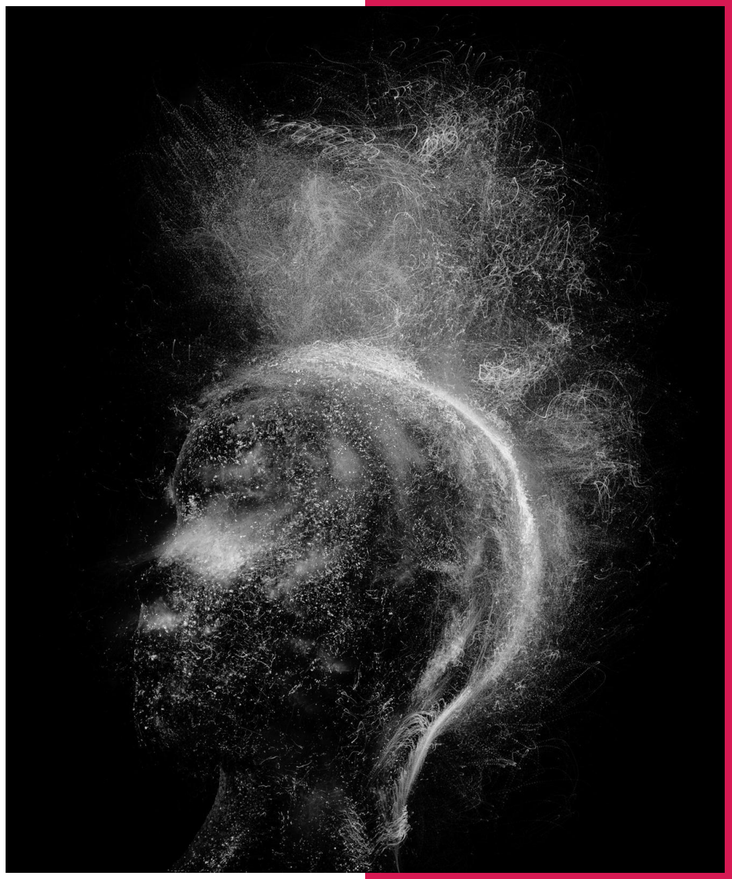
C’est une demande de plus en plus forte dans la société, et donc au sein des collaborateurs des entreprises, jusqu’à se retrouver dans les rapports RSE voire dans les objectifs ou missions que se fixent les entreprises. Autant l’anticiper, d’autant que des directives européennes pourraient à un moment fixer des seuils contraignants. Pour l’heure, le gouvernement en est au stade de l’incitation : il vient de publier un référentiel d’écoconception des services numériques (RGSN).

Cette nouvelle façon de voir l’IT doit imprégner notre usage des ressources. Qu’il s’agisse de notre façon de coder des applications ou des requêtes, ou de notre environnement matériel. L’équipement fourni aux collaborateurs évolue, avec notamment des ordinateurs légers et durables (en mode cloud-native, il suffit d’un navigateur Web pour tout), et des smartphones dont ils peuvent aussi se servir pour leur usage personnel (mode COPE : Corporate Owned, Personally Enabled). La durabilité, la réparabilité et le recyclage sont privilégiés, le cas échéant en conjonction avec des associations spécialisées.

Cela se retrouve aussi dans les efforts de réduction de l’usage des ressources cloud. C’est l’autre face du FinOps, dite GreenOps : l’approche mise en place pour éviter les dépenses inutiles sert (le plus souvent) à éviter de consommer inutilement des ressources. Elle se complète en outre de mesures spécifiques destinées à réduire l’empreinte environnementale (écoconception, choix des langages, etc.).

La Tech For Good, enfin, dépasse largement le cadre environnemental ; elle peut prendre plusieurs formes, toujours dans l’objectif de contribuer positivement à la société. C’est non seulement un enjeu RH de comprendre cela, mais aussi un devoir de s’en préoccuper.

PLATEFORMES



/04

PLATEFORMES

LES CONTRIBUTEURS DE CETTE SECTION

Seifeddin MANSRI
CTO Cloud Groupe,
SFEIR

Aurélien PELLETIER
Consulting Director Platform,
WENVISION

Feliks PLAWCZYK
Architecte, Consultant,
WENVISION

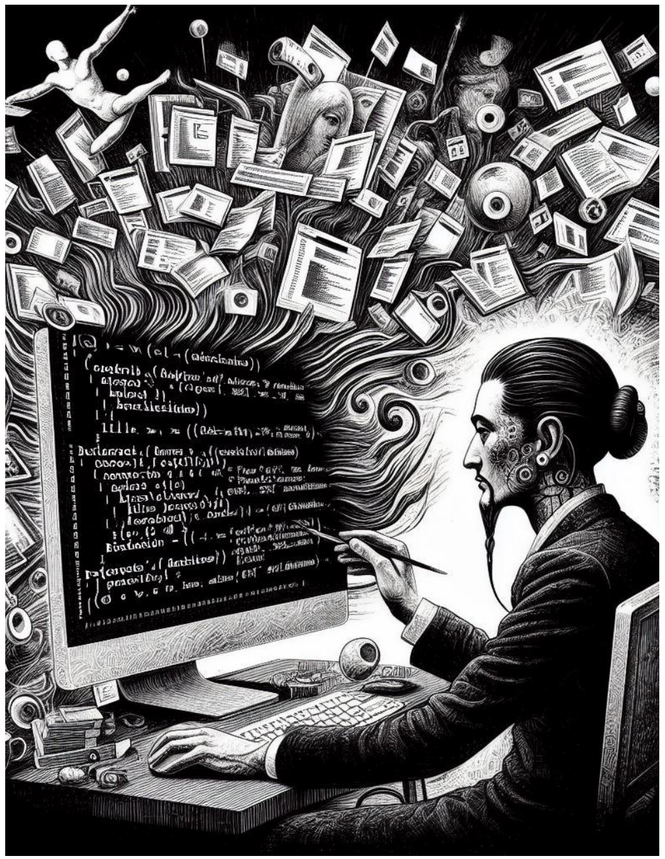
Tine KONDO
Senior Staff Engineer,
SFEIR

Olivier RAFAL
Consulting Director Strategy,
WENVISION

PLATEFORMES #Archi&Dev

L'architecture revient au centre du jeu

L'expérience développeur au cœur des plateformes



Dans le monde du développement logiciel, le succès de la création d'un produit repose sur l'efficacité des outils manipulés pour le construire et le déployer. On cherche toujours à réduire les frictions entre les équipes de développement et les infrastructures complexes, et les développeurs trouvent continuellement des pratiques modernes pour rendre l'ensemble du cycle de développement plus efficace.

Par exemple Microsoft, via ses solutions Github, consacre beaucoup d'énergie à la modernisation de ses outils pour répondre à la demande d'amélioration de la livraison des applications. Sa solution Copilot basée sur le LLM d'OpenAI est conçue pour améliorer l'efficacité du développeur. Des solutions comme Sourcegraph Cody, Tabnine ou bien Google Duet AI accroissent considérablement la productivité.

Des outils pour éviter le syndrome "ça marche en local"

De leur côté, les développeurs bénéficient de véritables plateformes intégrant toute la chaîne CI/CD (intégration continue, déploiement continu) pour industrialiser les déploiements. Des outils tels que Github Actions, Gitlab ou Azure DevOps permettent aux développeurs de configurer et de gérer plus facilement des déploiements complexes avec un minimum d'efforts. Un logiciel comme okteto, par exemple, permet de développer en local du code qui sera directement poussé dans Kubernetes. Le but est de fiabiliser et d'atteindre un certain niveau de reproductibilité.

En outre, ces outils contribuent à créer des environnements plus sécurisés et une meilleure évolutivité, permettant de lancer les applications plus rapidement et avec moins de problèmes. Et vu que les développeurs commencent à être jugés aussi sur les aspects GreenIT de ce qu'ils produisent, nous devrions voir arriver sous peu ce type de métrique "carbon footprint driven" (pour reprendre un vocable de Gitlab) et des aides au code prenant en compte cet aspect.

Pour améliorer la sécurité, les postes de travail des développeurs devraient devenir des terminaux passifs, où seule la partie interface homme-machine de l'environnement de développement sera présente sur le poste de travail. La partie métier de l'environnement de développement sera sur un serveur. Cela permettra de mettre en place la notion de Zero Trust sur le poste de travail des développeurs, car le code produit ne sera plus directement sur le poste de travail. Cela renforcera la sécurité en limitant l'accès aux données sensibles (voir Amazon CodeCatalyst, Google Cloud Workstation).

#Plateformes #Archi&Dev #Architecture #EDA

Event Driven Architecture : souplesse et réactivité

L'architecture pilotée par les événements (Event Driven Architecture - EDA) permet de développer des applications qui répondent en temps réel sans avoir à interroger continuellement le système pour obtenir de nouvelles informations. Cette architecture permet la surveillance d'événements métier qui peuvent déclencher des processus réactifs ; on parle donc aussi d'architecture réactive.

Plusieurs outils de type pub/sub (publish & subscribe, comme Google pub/sub ou AWS Kinesis) proposent un mécanisme simple, assurant la distribution de la donnée, tandis que des plateformes spécifiques proposent un environnement complet, incluant la possibilité de conserver les données transmises, voire les traiter ou en assurer la gouvernance. Dans ce domaine, l'implémentation de Kafka est très présente, qu'il s'agisse de la technologie open source ou d'offres managées comme celle de Confluent. Cependant, d'autres outils plus simples tels que RabbitMQ sont aussi toujours présents dans l'industrie. Et d'autres outils émergent également pour favoriser de nouveaux use cases, comme Apache Nifi.

De nombreux cas d'usage où le découplage est pertinent

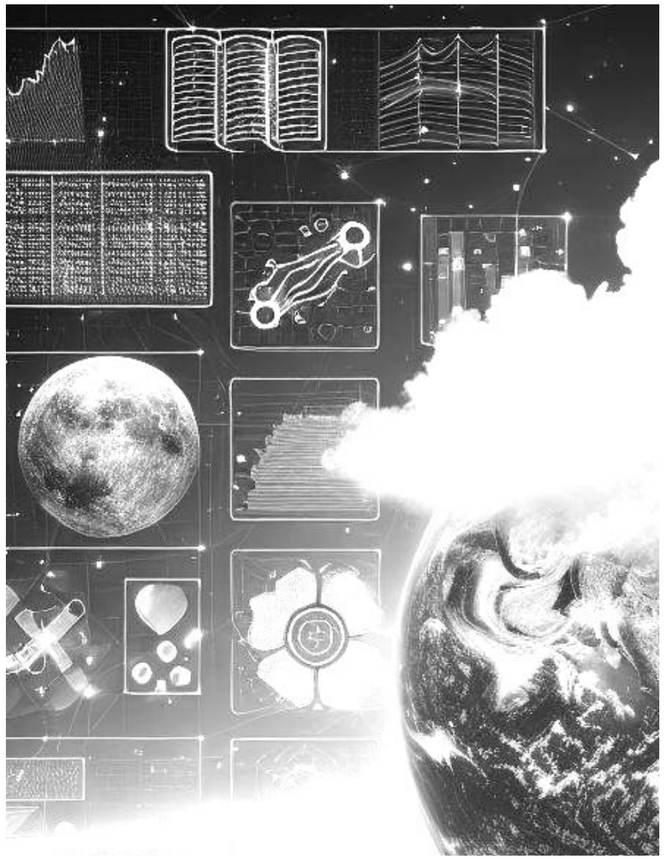
De ce fait, les outils et plateformes de diffusion d'événements au fil de l'eau ont une large palette d'utilisation, de la simple ingestion de données d'un outil transactionnel ou d'un capteur IoT vers une base ou une autre application (pour déclencher un processus, éditer une facture, etc.), par exemple, à du traitement complexe en temps réel, avec persistance des données à des fins d'audit ou de constitution d'un modèle de machine learning.

Ce type d'architecture peut contribuer à offrir de meilleures expériences aux utilisateurs tout en économisant le temps et les ressources associés à l'interrogation et à l'extraction de données. Elle offre également plus de souplesse au SI en favorisant le découplage entre des applications qui émettent ou consomment des événements, et donc davantage de résilience et des opportunité de scalabilité.

Les processus du SI ne sont plus figés, ils sont chorégraphiés.

#Plateformes #Archi&Dev #Architecture #MACH

MACH : pour l’agilité et l’innovation



MACH est un acronyme qui signifie Microservices, API-first, Cloud native et Headless. L’utilisation conjointe de ces approches est une tendance technologique qui a gagné en popularité ces dernières années et qui devrait continuer à gagner en importance, dans la mesure où cela pousse à adopter une réflexion et des architectures compatibles avec le futur. Penser MACH, c’est penser applications flexibles, évolutives et natives dans le cloud, utilisant les API comme principal moyen d'accéder à l'application et d'interagir avec elle. Cette approche permet une plus grande agilité et innovation, et est bien adaptée aux besoins des entreprises modernes.

MACH en détail

- **Les microservices** font référence à une architecture logicielle où l'application est construite comme une collection de petits services indépendants, plutôt que comme une base de code monolithique. Cette approche permet une plus grande flexibilité et évolutivité, car chaque service peut être développé et déployé indépendamment, et de nouveaux services peuvent être ajoutés ou supprimés selon les besoins.

  En revanche, cela introduit aussi de la complexité, il est donc important de bien cadrer et gouverner cette approche, en gardant en tête que l’important est surtout de modulariser l’applicatif. Le M de MACH pourrait ainsi se lire comme “Modules-based”.

- **API-first** signifie que l'application est conçue et construite autour des API, qu'elles soient REST ou GraphQL. Cette approche met l'accent sur les API comme principal moyen d'accéder et d'interagir avec l'application, plutôt que sur l'interface utilisateur.
- **Cloud native** fait référence aux applications qui sont conçues et construites pour fonctionner nativement dans le cloud. Cela signifie qu'elles sont construites à l'aide de technologies de type containers, serverless (fonctions déclenchables avec des événements, sans se soucier de la plateforme sous-jacente) ou autre capacités PaaS (Platform as a service) et qu'elles sont optimisées pour l'environnement cloud, ce qui les rend évolutives, résilientes et faciles à déployer et à gérer.
- **Headless** signifie que l'application n'a pas d'interface utilisateur, et qu'on y accède et la contrôle exclusivement par le biais d'API. Cette approche permet une plus grande flexibilité et personnalisation, car l'interface utilisateur peut être construite et modifiée indépendamment de l'application sous-jacente. Elle peut aussi, au besoin, être manipulée depuis une autre application, voire par un assistant LLM (agent basé sur de l’IA générative).

Les briques essentielles d’une plateforme GenAI

Une plateforme GenAI permet de répondre aux besoins en matière d'IA générative de la manière la plus efficace possible. Elle ouvre un monde de possibilités pour innover et améliorer les opérations de l’entreprise, en gardant ouvertes toutes les options possibles en matière de fournisseurs, de modèles, de cas d’usage ou encore d’intégration avec le SI et les processus de l’entreprise.

Les composantes a minima d’une plateforme GenAI :

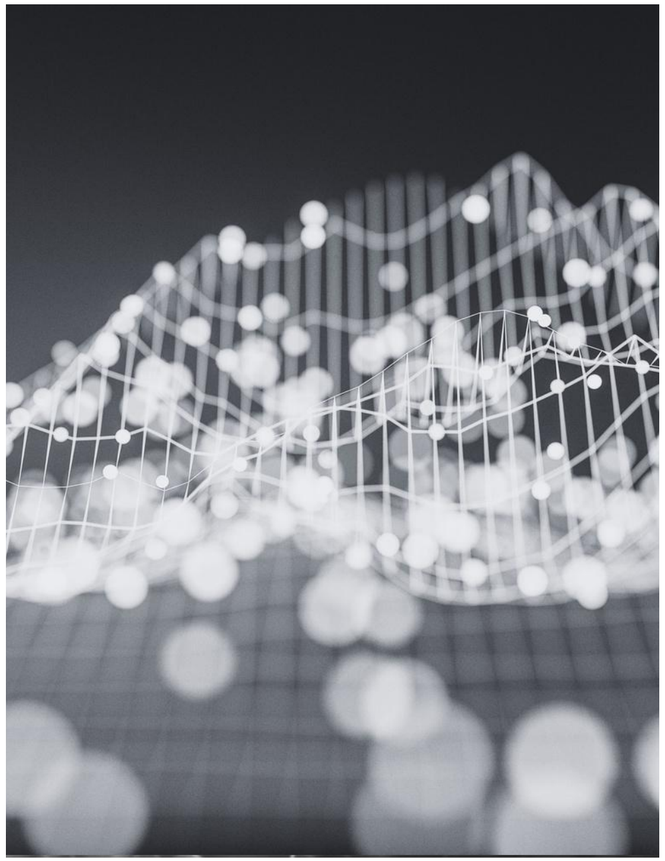
- **Chat intelligent :** avoir votre propre version de ChatGPT grâce à l'accès à un ou des LLM associés à une interface de chat.
- **Bibliothèque de connaissances :** pouvoir enrichir les connaissances du LLM avec vos documents, vos connaissances, votre contexte.
- **Boîte à outils de questions :** obtenir de meilleures réponses grâce à une collection de "formules magiques" (prompts) prêtes à l'emploi.
- **Orchestrateur :** utiliser un framework comme Langchain, qui s'assure que tous ces outils fonctionnent bien ensemble.
- **Gardien éthique :** découvrir les usages qui sont faits de l’IA dans l’entreprise et en garantir un usage éthique.
- **API :** offrir ces services sous forme d’API pour aller au-delà de l’interface de chat et intégrer ces technologies dans votre SI.

PLATEFORMES #SIcloud

Le SI Cloud devient une réalité

#Plateformes #SIcloud #CloudNative

Cloud Native : fiabilité, évolutivité et agilité



L'adoption d'une stratégie Cloud Native simplifie le développement et la maintenance des applications, tout en les rendant plus résilientes, agiles et élastiques. C'est le C de l'acronyme "Mach" (micro-services, API, cloud native et headless) : il s'agit de bâtir le SI sur une infrastructure capable d'évoluer en fonction des impératifs business et des optimisations technologiques.

Un SI Cloud Native est une combinaison d'applications SaaS, de services managés et d'applications ou micro-services maison conteneurisés, déployés dans le cloud. Ce modèle offre une fiabilité, une évolutivité et une agilité constantes. Avec l'automatisation, les tâches sont souvent accomplies plus rapidement et avec plus de précision en déchargeant les équipes informatiques des tâches routinières et sujettes aux erreurs. Les opérations d'administration, de maintenance et de sauvegarde sont prises en charge par la plateforme, les équipes pouvant se concentrer sur des tâches à plus forte valeur ajoutée comme l'orchestration.

SaaS, conteneurisation ou services managés : des choix stratégiques qui dépendent du contexte

En l'absence d'une offre SaaS, le choix entre un service managé et la conteneurisation dépend des besoins de l'entreprise. Les deux approches sont parfaitement valables, et peuvent être mixées en fonction des workloads considérés. En adoptant les services managés, les entreprises créent une certaine adhérence au fournisseur, mais gagnent énormément sur le déploiement et la maintenance, libérant ainsi du temps et des ressources.

La conteneurisation offre une plus grande flexibilité et réduit l'adhérence des workloads à un cloud provider, mais elle nécessite plus de compétences et de ressources. En effet, la relation présente avec un fournisseur n'augure jamais de la relation future. En réduisant son adhérence à un fournisseur de cloud, l'entreprise s'achète la liberté future.

#Plateformes #SIcloud #Sécurité #ZeroTrust

Zero Trust Network : sécurité et flexibilité

Le modèle Zero Trust Network prend pour principe que le réseau de l'entreprise ne doit pas être un élément considéré comme sécurisé. De fait, les SI s'ouvrent largement : applications SaaS, hybridation cloud et on premises, utilisateurs à distance, wifi plus ou moins ouvert, plateformes pour les partenaires, espaces clients...

Dans ce contexte, où les contours du SI deviennent flous, il n'est plus possible de se contenter d'une sécurité périmétrique. L'approche zero trust permet de s'assurer que toutes les demandes d'accès à des applications ou des données doivent être vérifiées (authentifiées et autorisées), que ce soit au sein d'une organisation, d'un partenaire ou d'un client : ce n'est pas parce qu'un utilisateur / une machine est connecté(e) au réseau de l'entreprise qu'il/elle doit être considéré(e) comme fiable.

Internet devient le backbone des entreprises modernes

Le concept de sécurité périmétrique est réduit à son minimum : le cluster de machines. L'hypothèse de base est de dire que le réseau de l'entreprise est aussi sécurisé qu'internet. Certaines entreprises vont même jusqu'à dire que le réseau de l'entreprise, le backbone même, est internet. Ce postulat, très puissant, permet aux entreprises qui l'adoptent de complètement se délester des coûts de mise en œuvre de l'infrastructure réseau et amène une très grande flexibilité pour mettre en œuvre un modèle d'entreprise plateforme, au centre de son écosystème, ainsi que des solutions qui permettront aux salariés de rester connectés en permanence au SI.

Ce modèle implémente la 'security by design' ou la sécurité en profondeur qui est incorporée dans l'ensemble des briques du SI faisant intégralement partie de son ADN.

Le Zero Trust Network peut être aussi considéré comme une culture ou un état d'esprit qui doit être un réflexe naturel chez l'ensemble des salariés de l'entreprise - ce qui nécessite une phase d'acculturation et de sensibilisation pour ces derniers.

#Plateformes #SIcloud #Innovation

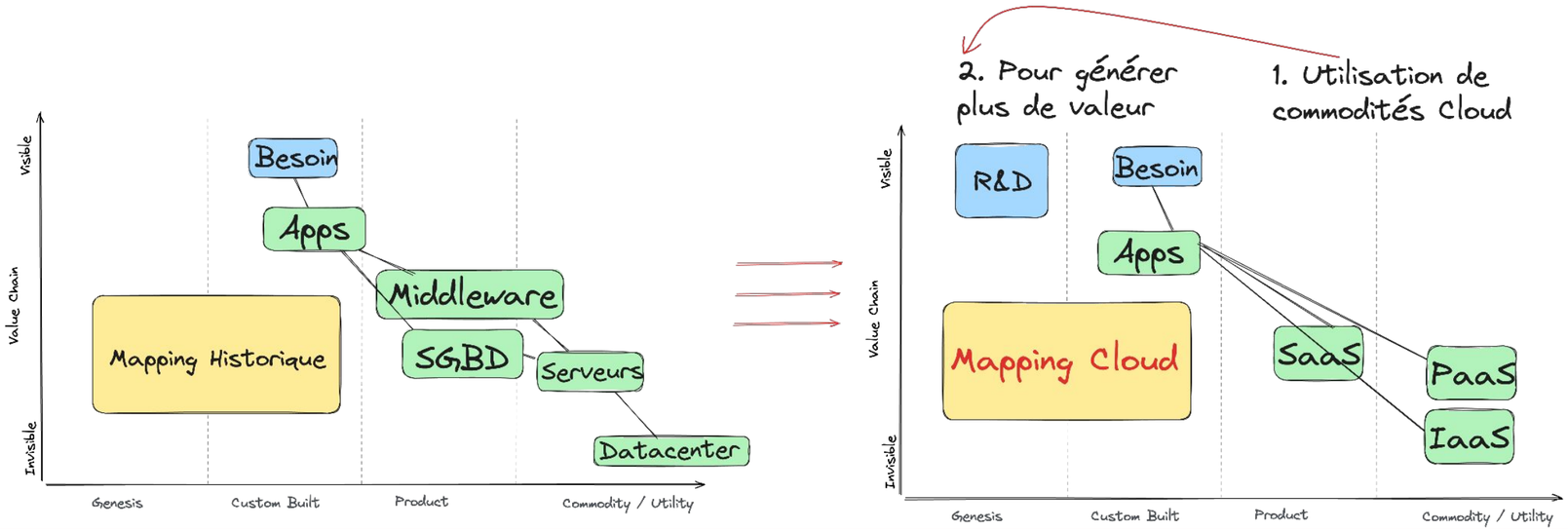
Les services du cloud deviennent des commodités : innovons !

Il y a quelques années, les services offerts par les cloud providers publics devaient faire l'objet, comme toute innovation, d'une appropriation (tant culturelle que technique) par les équipes techniques et métiers. Ces éléments pouvaient être considérés, en les observant par le prisme d'une cartographie de Wardley (décrivant la chaîne de valeur), comme appartement à la catégorie "genesis", alors que les serveurs typiquement présents dans un datacenter faisaient partie de la catégorie "commodities".

Cette évolution doit être prise en compte dans les architectures applicatives définies en 2024. Les services de types PaaS (et IaaS dans une certaine mesure) doivent être considérés comme des "commodities" et utilisés comme tels afin de contribuer à générer de la valeur pour les clients et utilisateurs. À de rares exceptions près, ceux-ci présentent toutes les garanties requises pour un usage en entreprise, tant en termes de gouvernance, de capacité et de sécurité.

Les entreprises doivent franchir ce pas et ne plus s'engager dans la construction "in-house" de tels services. Les cloud providers eux-mêmes l'ont bien compris, en fournissant des produits construits sur ces briques de base à leurs clients.

En 2024, générer de la valeur passera par la construction de produits innovants basés sur ces commodités matures, plutôt que construire ces commodités via des programmes internes coûteux et consommateurs de temps et de ressources.



#Plateformes #SIcloud #FinOps

Le FinOps boosté par la situation économique

Créer des plateformes et des produits dans le cloud est un investissement générant de la valeur pour l'entreprise. Après des décennies où l'IT était uniquement vue comme un centre de coûts, cette vision orientée produits rebat les cartes, d'autant qu'avec le cloud, il devient possible de connaître très précisément le coût de chaque service, à la seconde près. Et de le ventiler en fonction des produits business - dès lors qu'on aura bien pris la précaution de labelliser les ressources cloud.

Ce plan d'étiquetage des ressources est la pierre angulaire de la démarche FinOps, contraction d'opérations et de finances. Les responsables FinOps animent, orchestrent cette démarche, acculturent les utilisateurs (IT, métiers et financiers), diffusent les bonnes pratiques, et enfin rassemblent, extraient et préparent les données pour que :

- L'IT puisse optimiser son usage, réduire son utilisation des ressources au strict nécessaire, avec des impacts positifs autant pour les finances que pour l'environnement.
- Les directions puissent analyser les performances et prendre des décisions stratégiques pour l'entreprise.

Entre l'augmentation des dépenses prévisibles dans le cloud (davantage d'utilisateurs, périmètre élargi, ajout de fonctionnalités...) et la nécessité de contrôler ses investissements au vu d'une situation économique incertaine, le FinOps devrait venir tout en haut des préoccupations.

Faire des dépenses cloud un signe de croissance et de réussite business

Une stratégie FinOps efficace ne se limite pas à réduire les coûts, mais vise à optimiser l'utilisation des ressources cloud dans une optique de pilotage par la valeur : il n'y a pas de souci intrinsèque à investir dans le cloud dès lors qu'on s'assurer de créer de la valeur métier (chiffre d'affaires direct, amélioration des temps de réponse, ajout de fonctionnalité optimisant la rétention client, etc.). Dans une situation saine, l'augmentation des coûts cloud est un signe de croissance et de réussite commerciale.

#Plateformes #SIcloud #PrivateCloud

Le cloud privé et déconnecté pour quelques cas d’usage

Les débats sur le cloud souverain et le cloud de confiance (cf. notre section #culture sur la souveraineté) ont mis en évidence des besoins, et finalement donné naissance à une nouvelle forme de cloud, où les fournisseurs cloud deviennent fournisseurs d'infrastructures de cloud sans en assurer le run.

Les 3 fournisseurs de cloud majeurs ont scellé des alliances stratégiques avec des acteurs locaux pour créer des offres répondant à la notion de cloud de confiance :

- S3NS, la co-entreprise managée par Thalès, qui s’appuie sur les technologies de Google Cloud.
- Bleu, née de l'alliance entre Orange, Capgemini et Microsoft.
- Le cloud de confiance d’Atos et AWS, dont le nom n’est toujours pas défini.

Ces offres déconnectées du réseau principal du fournisseur de cloud, opérées par des entreprises françaises ou européennes, constituent une solution intéressante pour des organismes publics et répondent également aux dilemmes de certaines entreprises de type OIV (Opérateur d’importance vitale) ou OSE (Opérateur de service essentiel), et/ou faisant partie de secteurs très règlementés (banques, assurances, santé…), qui souhaitent bénéficier des capacités d’innovation et de prospérité du cloud telles que des services managés (et allant donc au-delà du simple hébergement de machines virtuelles) tout en protégeant les données sensibles et garantissant le respect des contraintes règlementaires applicables.

Le label SecNumCloud ne peut pas s’appliquer à toutes les technologies Cloud

Toutes ces initiatives ont pour objectif décrocher le label ‘SecNumCloud’ de l’ANSSI qui impose aux organismes publics, administrations, opérateurs d’importance vitale et opérateurs de services essentiels d’utiliser des clouds européens certifiés ‘SecNumCloud’.

Toutefois, il est à noter que tous les services présents dans l’offre publique du fournisseur de cloud ne seront pas nécessairement présents dans les offres dites de confiance pour des raisons technologiques évidentes (lorsque, par nature, l’offre ne peut satisfaire aux contraintes imposées par le label).

#Plateformes #SIcloud #Sustainability

Le cloud au cœur des questions d’environnement et d’énergie

Les notions de “sustainability”, d’impact carbone et d’approvisionnement énergétique deviennent cruciales (cf. aussi à ce sujet notre section #culture). Une récente enquête PAC/Numeum indique que dans les trois quarts des appels d’offres, les fournisseurs du numérique doivent expliquer et démontrer leurs actions de type RSE.

A cela s’ajoute la nécessité d’assurer la résilience de son IT en cas de coupures de courant. Les fournisseurs de cloud ont anticipé ces problématiques et prennent les devants pour apporter des solutions concrètes.

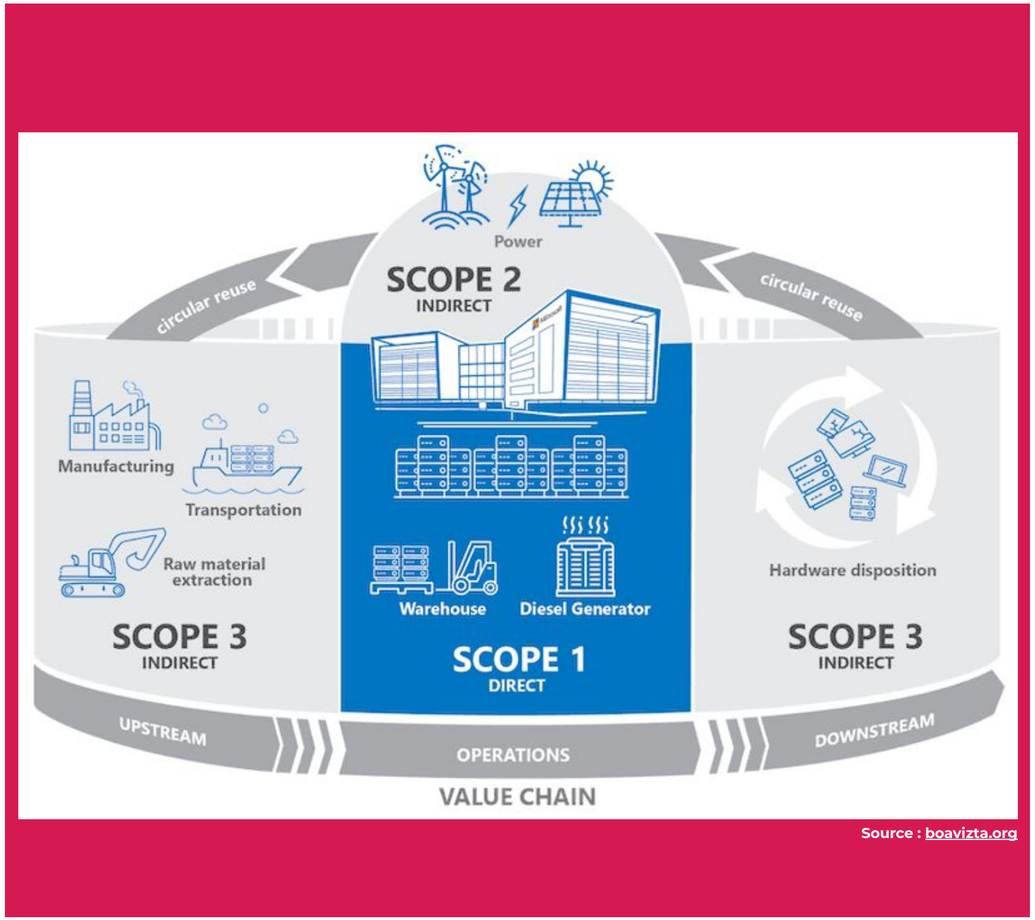
Si l’usage du cloud est le plus souvent bien plus efficient du point de vue de la consommation de ressources et d’énergie (par rapport à de la production IT en interne), son poids environnemental doit être pris en compte dans les bilans.

Et depuis le 1er janvier 2023, la réalisation d'un bilan carbone complet (scopes 1, 2 et 3) est obligatoire pour les entreprises de plus de 500 salariés en métropole et pour les entreprises de plus de 250 salariés en outre-mer.

Il est ainsi primordial de connaître les différents scopes et surtout leur signification pour l’entreprise :

- Scope 1 : émissions directes de gaz à effet de serre comme celles liées au chauffage ou à l’usage des véhicules de l’entreprise.
- Scope 2 : émissions indirectes liées principalement à l'électricité et son mode de production.
- Scope 3 : toutes les autres émissions indirectes liées notamment aux activités de l’entreprises comme les chaînes d’approvisionnement, les fournisseurs ou partenaires.

Sans surprise, le scope 3 est le plus compliqué à déterminer (d’autant qu’aucun fournisseur de cloud ne communique publiquement dessus) mais le négliger donne une vision incomplète de l’impact environnemental de l’entreprise.



Source : boavizta.org

#Plateformes #SIcloud #Sécurité

La sécurité “as code” devient obligatoire

Dans un monde de plus en plus hostile, le cloud est une des solutions pour augmenter le niveau minimum de sécurité des entreprises. D'une part, parce que la sécurité est au cœur des préoccupations et des investissements des fournisseurs de cloud - c’est primordial pour leur business model : leur survie en dépend.

D’autre part, parce que les offres de sécurité managées sur le cloud pallient le manque de talents en cybersécurité, en externalisant une partie de la stratégie de sécurité et en déléguant les couches les plus basses aux fournisseurs de cloud dans un modèle de responsabilité partagée.

Au-delà du respect des bonnes pratiques d'une architecture cloud sécurisée, cette attention à la sécurité devra se matérialiser sous forme de code. Tout comme une infrastructure cloud est scriptée (“infra as code, IaC”), la sécurité devient aussi “as code”.

Les règles de sécurité intégrées à l’infra as code

On retrouve ainsi :

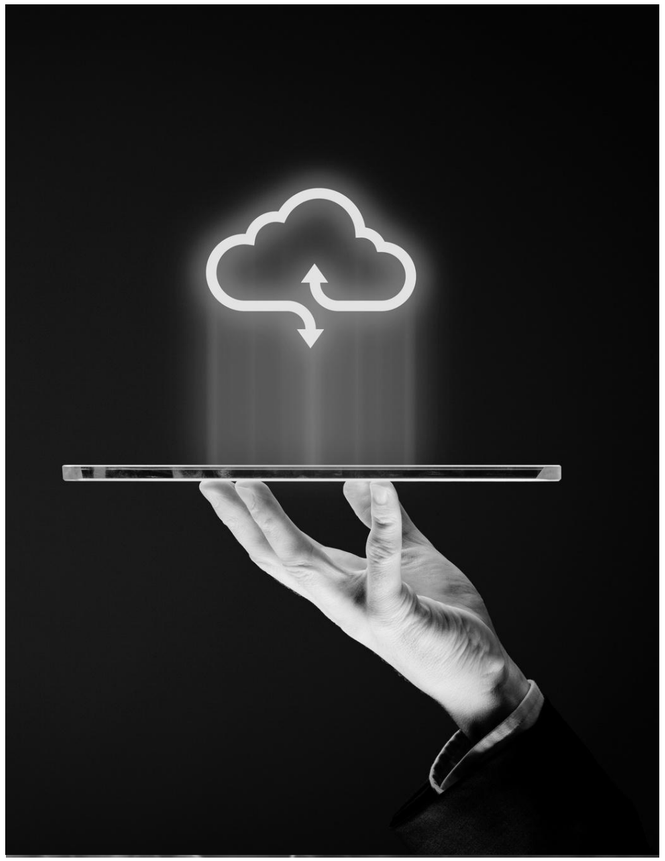
- **La politique de sécurité as code** : l'ensemble des règles de sécurité communes à l'ensemble des projets de l'organisation, dont il ne sera pas possible de s'écarter. Par exemple, l’interdiction globale de rendre public n'importe quel bucket appartenant à l'organisation (via Google Organization Policies, AWS Security Control Policies ou encore Azure Security Policies).
- **Les règles de configuration des ressources d’infrastructure cloud as code**, qui inspectent les scripts IaC : ce sont les règles de bonne configuration de chaque produit cloud qui sont évaluées avant le déploiement et le stoppent si une non conformité a été détectée (exemples d’outils : Checkov, KICS, Spectral, tfsec...).
- **La détection de dérives** (“drift”) par rapport aux configurations initiales, en cas de changements manuels.

Dans ce domaine, les travaux de la Cloud Native Computing Foundation sur OPA, Open Policy Agent, seront intéressants à suivre : OPA est un langage déclaratif permettant de spécifier une “policy as code” à appliquer dans les microservices, Kubernetes, pipelines CI/CD, etc.

Enfin, les plateformes SecOps modernes s’appuient de plus en plus sur des règles de détection codées à l’image de Chronicle SIEM avec YARA-L, un langage de détection des menaces modernes.

#Plateformes #SIcloud #PlatformEngineering

Le Cloud Platform Engineering prend son envol



Le cloud platform engineering est une discipline qui consiste à concevoir, développer et mettre à disposition des infrastructure en self service dans le cloud. Ces plateformes fournissent un ensemble de services et d'outils permettant aux développeurs de travailler de manière efficace et agile.

Une équipe d'ingénierie centrale est souvent en charge d'une telle plateforme. Elle est responsable de la construction d'un catalogue de produits prêts à l'emploi à travers des templates qui automatisent la création des composants effectifs d'infrastructure. Des standards partagés et acceptés par les différentes équipes sont nécessaires pour une adoption efficace de la plateforme. Les mots d'ordre sont la mutualisation et la réutilisation pour éviter de réinventer la roue par chaque équipe dans son coin.

Le cloud platform engineering comprend principalement deux parties :

- Le core platform engineering, avec tous les outils et technologies nécessaires au provisioning des composants d'infrastructure (internal developer platform).
- Le portail de consommation des services fournis (internal developer portal).

L'outillage commence à prendre corps

Des outils techniques commencent à émerger pour implémenter les principes théoriques du cloud platform engineering et proposer un cadre de développement.

Backstage fait partie des outils à suivre. Créé initialement par Spotify, il est actuellement en incubation à la CNCF (Cloud Native Computing Foundation). Il s'agit d'un produit open source permettant de construire un portail pour les développeurs, dans lequel ils peuvent se servir, en s'appuyant sur des composants natifs tels que le catalogue, les templates et la documentation technique.

La combinaison Backstage et Terraform s'avère puissante pour les opérations, magique pour les développeurs.

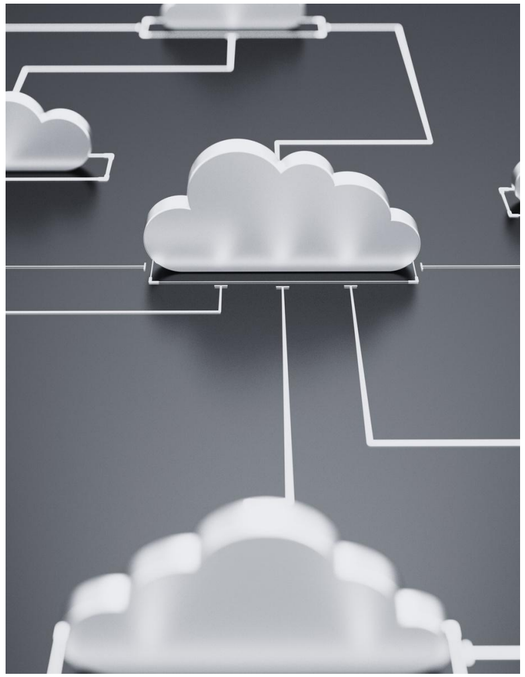
#Plateformes #SIcloud #edge

Edge Cloud : un cloud à portée de main

L'Edge Cloud est une architecture informatique qui consiste à placer des ressources cloud computing à proximité des sources de données, c'est-à-dire à la périphérie du réseau. Cela permet de réduire la latence et d'améliorer les performances des applications, en particulier celles qui nécessitent un traitement en temps réel des données ou les sites avec une connectivité Internet faible ou quasi-inexistante.

L'Edge Cloud est une technologie émergente qui présente un grand potentiel pour de nombreux secteurs, tels que la vente au détail, la santé, l'industrie ou encore les transports.

Prenons l'exemple des retailers qui continuent d'investir dans la technologie : ils ont de plus en plus besoin d'exécuter des logiciels, tels que les systèmes de point de vente, ainsi que de nouvelles capacités d'IA et d'apprentissage automatique au sein de ces systèmes, avec une connectivité Internet à faible latence ou intermittente dans les emplacements physiques. De plus, de nombreux retailers souhaitent étendre les avantages des applications cloud dans leurs magasins à grande échelle avec des niveaux de sécurité des données élevés.



Assurer la continuité des traitements, même sans connexion

Ils peuvent ainsi économiser des ressources informatiques précieuses grâce à une poignée de serveurs de format réduit gérés par un fournisseur de Cloud et qui peuvent être installés facilement dans n'importe quel magasin. Grâce à cette extension de plateforme dans la salle du fond d'un magasin, les détaillants peuvent effectuer toutes sortes de tâches, de la collecte d'analyses de magasin complètes à la simplification des opérations critiques des magasins.

Même constat pour les industriels, qui ont besoin d'assurer la continuité de service dans les usines même en cas d'interruption de la connectivité ou d'indisponibilité rare mais possible du cloud public. Ils auront aussi la possibilité de tirer profit des technologies, peu utilisées dans le monde industriel, comme la conteneurisation ou l'IA.

PRODUITS

Tendances Tech | 2024

/05

PRODUITS

LES CONTRIBUTEURS DE CETTE SECTION

ILS ET ELLES ONT ÉCRIT LES TECH TRENDS 2024

Sandrine BOITEAU
Consulting Director Orga & Product,
WENVISION

Lucas DAHAN
Product Strategy Consultant, Data & IA Expert,
WENVISION

Aude DEFRETIERE
Consultante Product Strategy,
WENVISION

Romain MAZUIR
Head of Product Strategy,
WENVISION

Olivier RAFAL
Consulting Director Strategy,
WENVISION

Giulia TAIBI
Product Strategy Consultant, Digital Expert,
WENVISION

PRODUITS
#ProcessusProduits

Tendances Tech | 2024

Les clés du Product Management

#Produits #ProcessusProduits #PLM

En finir avec les big bangs IT : la gestion du cycle de vie des produits

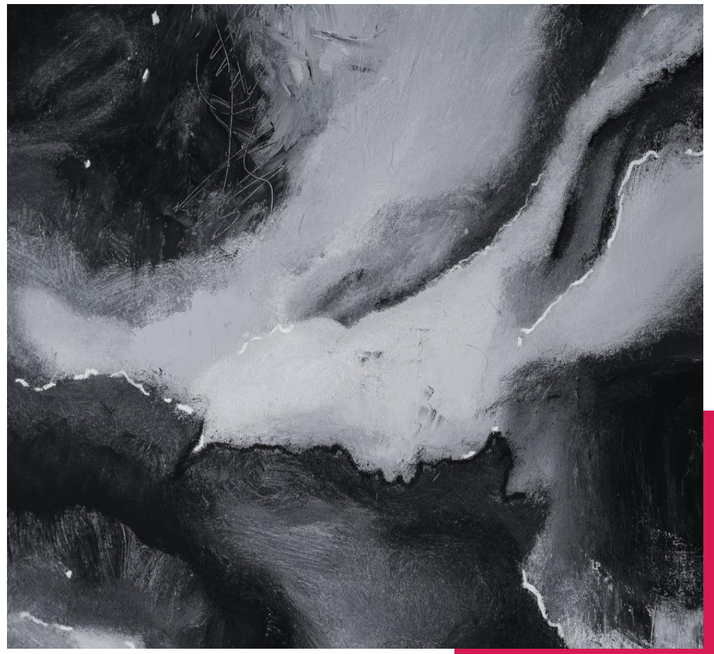
L'environnement IT d'une entreprise est complexe et en constante évolution. Les nombreux défis comme la gestion efficace des ressources, l'adaptation aux changements technologiques rapides et la maximisation du retour sur investissement accélèrent le rythme des transitions technologiques. Ces transitions ralentissent la croissance de l'entreprise en gelant les évolutions des actifs IT nécessaires au soutien des objectifs business. Changer le système de gestion de ces actifs pour réduire les effets de ruptures devient clé : c'est ce qu'offre le cadre du Product Lifecycle Management (PLM).

Le PLM, pour gérer les actifs informatiques comme des produits

Le PLM dans l'IT implique de gérer les actifs informatiques comme des produits tout au long de leur cycle de vie, depuis la conception jusqu'à la mise hors service. Cette perspective change la manière dont les actifs IT sont perçus et gérés, en les considérant non plus comme des coûts fixes, mais comme des actifs stratégiques qui évoluent et apportent de la valeur à l'entreprise.

Les bénéfices du PLM sont multiples. Tout d'abord, cela permet un alignement stratégique plus précis, où les actifs IT sont gérés en synergie avec les objectifs business, garantissant ainsi que les technologies contribuent activement aux plans de l'entreprise. Cette méthode améliore également la gestion des ressources, optimisant la planification et l'allocation des investissements IT pour une utilisation efficace. De plus, elle renforce la réactivité et la flexibilité de l'entreprise, la rendant plus adaptée à répondre rapidement aux changements technologiques et aux variations du marché.

Le PLM veut limiter les effets de rupture, ralentissant l'atteinte des objectifs stratégiques, en proposant un cadre de gestion favorisant une transition proactive des actifs IT.



#Produits #ProcessusProduits #ValueStream

Le Value Stream Management s’opère en Value Stream Analytics

En 2024, la gestion des flux de valeur (Value Stream Management, VSM) évolue pour devenir un élément central dans la pilotage stratégique des entreprises. Cette approche, initialement issue des pratiques industrielles du lean, se focalise désormais sur la nécessité de considérer les flux de valeur comme des produits à part entière, dotés de leurs indicateurs clés alignés sur les objectifs stratégiques et le retour sur investissement.

La pratique moderne du VSM va au-delà de la simple identification et cartographie des différentes étapes de création d’un produit (Value Stream Identification et Mapping). Elle s'oriente vers une analyse approfondie et continue de la création de valeur à chaque étape du processus. Cette évolution s'inscrit dans le cadre de l'utilisation de cadres d'agilité à grande échelle, tels que SAFe, favorisant ainsi des flux continus de livraison et de développement.

Au-delà du lean : la recherche du ROI des flux

Le concept de Value Stream Analytics prend de l'importance, offrant aux entreprises des outils pour mesurer précisément la valeur générée par ces flux en termes de résultats commerciaux. Cette évolution est particulièrement pertinente dans un contexte économique incertain, où maximiser l'efficacité et le retour sur investissement devient crucial.

L'objectif n'est plus seulement de produire un logiciel ou un produit, mais de créer un flux de valeur intégré et optimisé qui soutient l'ensemble de l'entreprise. Cette tendance va impacter les organisations et le mode de gouvernance des investissements..

#Produits #PortfolioManagement #Agilité

De l’agilité à la gestion des portefeuilles produits

L’agilité et le portfolio management sont des concepts séduisants, qui pourront s’avérer efficaces ou non selon la manière dont ils sont implémentés.

La réorganisation systématique des équipes IT/techno avec des remèdes clés en main ne résoudra pas vos problèmes. Loin d’une solution miracle, la mise en œuvre d’une gestion des portefeuilles produits est un exercice qui demande de changer l’organisation en profondeur. Avec un potentiel énorme pour le business, qui verra alors les bénéfices d’une organisation orientée produits.

La mise en place d'équipes agiles, tout comme la nomination de Product Managers, s'avère insuffisante sans une vision globale et des structures organisationnelles solides. Les entreprises doivent penser global si elles veulent voir émerger une nouvelle ère d'agilité et de création de valeur. Les Product Managers ne doivent plus être laissés seuls face à des choix impossibles, mais soutenus par une vision stratégique claire et une collaboration fluide entre toutes les parties prenantes de l'entreprise. C'est le moment d'embrasser une approche globale pour façonner l'avenir prospère de l'entreprise : la gestion des portefeuilles produits.

Les bénéfices du portfolio management

- Un alignement stratégique

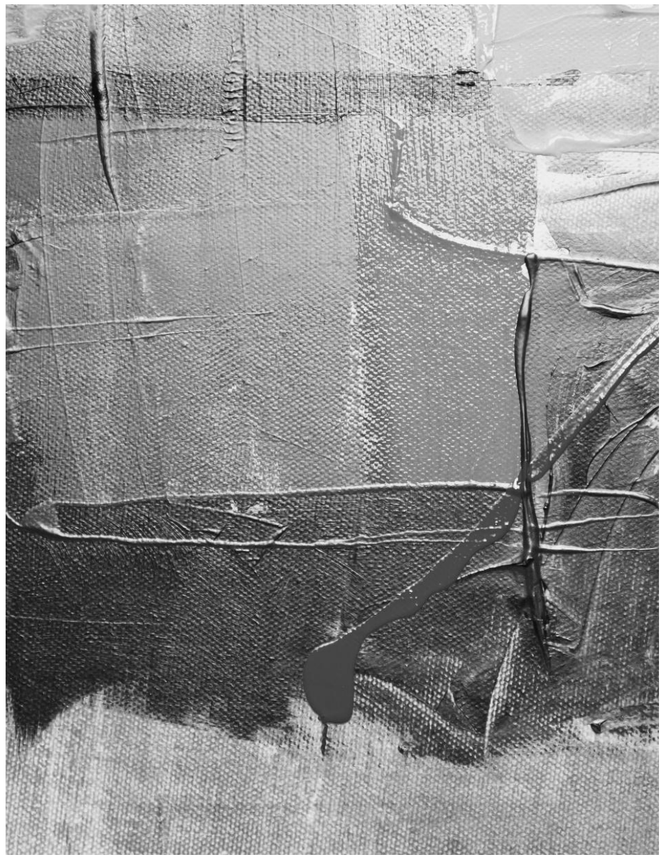
Alors que la gestion de projet se concentre sur la réalisation d'objectifs spécifiques, la gestion de portefeuille produit aligne chaque initiative sur une vision stratégique plus large. Cette synchronisation stratégique est cruciale pour garantir que chaque initiative contribue de manière significative à la réalisation des objectifs à long terme de l'entreprise. Les différents produits vont donc être régulièrement évalués à travers l’exercice de portefeuille pour vérifier leur adéquation avec les objectifs stratégiques.

- Une répartition équilibrée des efforts

Un portefeuille bien géré ne se contente pas de regrouper des produits. En intégrant des produits à différentes étapes de maturité, l'entreprise crée une dynamique qui assure une croissance continue. On peut ainsi réserver une bande passante pour les “vaches à lait”, du temps pour les innovations ou produits à fort potentiel, et un peu de run pour les produits sur le déclin. La gestion de portefeuille produit cherche le mix idéal. Un équilibre entre les produits matures générant des revenus stables et les produits innovants prêts à conquérir de nouveaux marchés crée une symphonie de croissance.

#Produits #ProcessusProduits #PortfolioManagement

Le Product Management encadré par du PPM



Dans des entreprises qui disposent d'un catalogue de produits étendu, le PPM s'avère un bon moyen d'optimiser ses efforts afin de maximiser la création de valeur. **Deux éléments vont être importants dans le découpage des portefeuilles** afin d'atteindre ce but.

1. Le choix du critère de découpage

Un "portefeuille produit" est l'agrégation de tous les actifs liés aux produits d'une entreprise, qu'ils soient déjà sur le marché, en cours de développement ou en phase de planification. Cet ensemble peut comprendre une gamme variée de produits, de projets de développement, d'initiatives d'innovation, ou encore d'activités de maintenance et d'amélioration.

La gestion d'un portefeuille produit implique de prendre en compte l'ensemble de ces actifs et de coordonner leur gestion de manière stratégique. Le choix de la méthode de découpage dépendra des besoins spécifiques de l'entreprise, de ses objectifs stratégiques et de sa capacité à gérer les portefeuilles de manière efficace. Souvent, une combinaison de plusieurs de ces méthodes est utilisée pour répondre à différentes dimensions de la gestion des portefeuilles produits. Quelques exemples de critères de découpage : Business Units, Domaine produit, objectif stratégique, segmentation client...

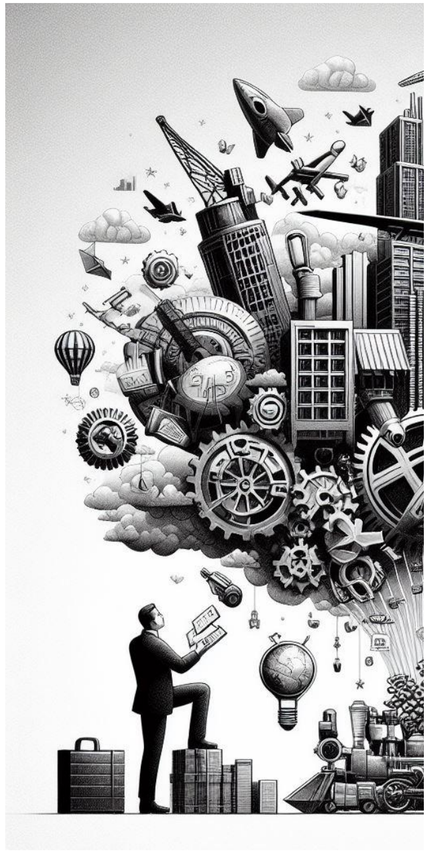
2. L'équilibre des produits au sein du portefeuille

La gestion de portefeuille, qu'elle soit axée sur les produits ou les actifs financiers, partage un principe fondamental : la nécessité de diversification pour atténuer les risques. A l'image d'un investisseur qui diversifie son portefeuille d'actions pour réduire la volatilité, un responsable produit doit s'efforcer de créer un portefeuille équilibré pour minimiser les risques liés à des décisions stratégiques isolées.

Les fluctuations du marché, les évolutions technologiques et les changements dans le comportement des consommateurs sont autant de facteurs de risque qui peuvent influencer la performance d'un produit. En comprenant ces risques et en les intégrant dans la stratégie de gestion du portefeuille produit, les entreprises peuvent mieux anticiper les défis potentiels et ajuster leur positionnement en conséquence. Ces risques vont évoluer fortement en fonction de la maturité du produit dans son cycle d' évolution (Life Cycle Product).

#Produits #ProcessusProduits #PM&PO

Le duo gagnant Product Manager et Product Owner



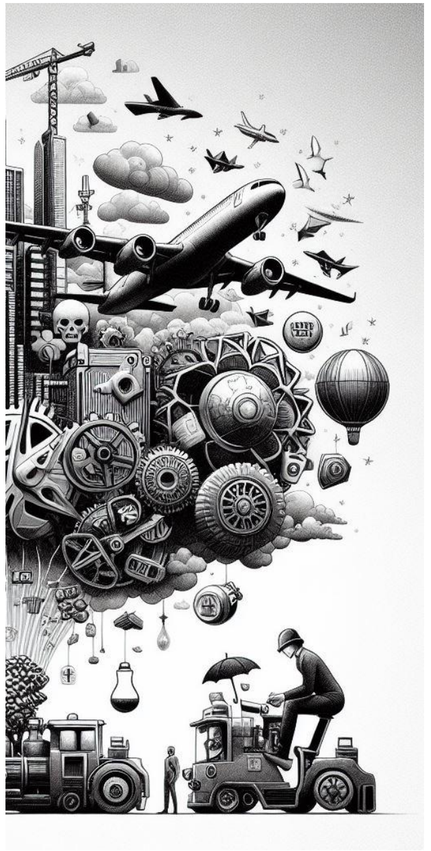
Les rôles de Product Owner (PO) et de Project Manager (PM) sont souvent sujets à débat dans le monde de la gestion de produit. Néanmoins, la synergie de ces deux rôles dans la structuration et le déploiement d'un produit est évidente ; c'est le prélude d'une stratégie gagnante.

- Le PO, en se focalisant sur le delivery de la solution, l'interface avec les utilisateurs, ainsi que sur la gestion efficace du backlog, joue un rôle crucial dans l'optimisation du processus de développement du produit. Cette focalisation permet de s'assurer que le produit reste aligné avec les besoins et les attentes des utilisateurs finaux, garantissant ainsi sa pertinence sur le marché.
- Le PM, avec son regard tourné vers la stratégie globale, la gestion des parties prenantes, ainsi que la veille sur les tendances et les évolutions du marché, apporte de nouvelles perspectives et orientations stratégiques au produit. Cette approche plus holistique aide à aligner le développement du produit avec les objectifs à long terme de l'entreprise et les réalités du marché.

Si l'un des deux joue les 2 rôles, l'équilibre est rompu

Leur association permet non seulement d'optimiser le time to market pour les différentes releases, mais assure également une continuité et une cohérence dans les activités opérationnelles et stratégiques. Cette collaboration permet au Product Manager de se concentrer sur des aspects vitaux, tels que les interactions nécessaires avec les équipes Commerce, Marketing et Communication, ainsi que les équipes Support pour assurer une diffusion efficace et une adoption réussie du produit sur son marché.

A contrario, lorsque le PM assume les fonctions du PO, cela introduit des ruptures dans le processus. Le PM doit alors jongler entre la définition et l'application de la vision du produit, ce qui génère de l'inertie dans l'exécution de la stratégie.



#Produits #ProcessusProduits #Discovery #IA

L’importance du discovery pour la construction de produits IA

La phase de discovery est une étape préliminaire cruciale dans le processus de développement de produits, essentielle pour comprendre les besoins des utilisateurs, les tendances du marché et les défis techniques. Elle vise à clarifier la stratégie qui permettra d’aligner les besoins des clients, la vision de l’entreprise, les possibilités et contraintes techniques, ainsi que les enjeux commerciaux avant de commencer le développement des premières versions du produit.

Cette étape de discovery offre de nombreux avantages. Elle réduit les risques en identifiant les problèmes potentiels à l'avance, optimise l'allocation des ressources et améliore la pertinence des solutions choisies en les évaluant préalablement.

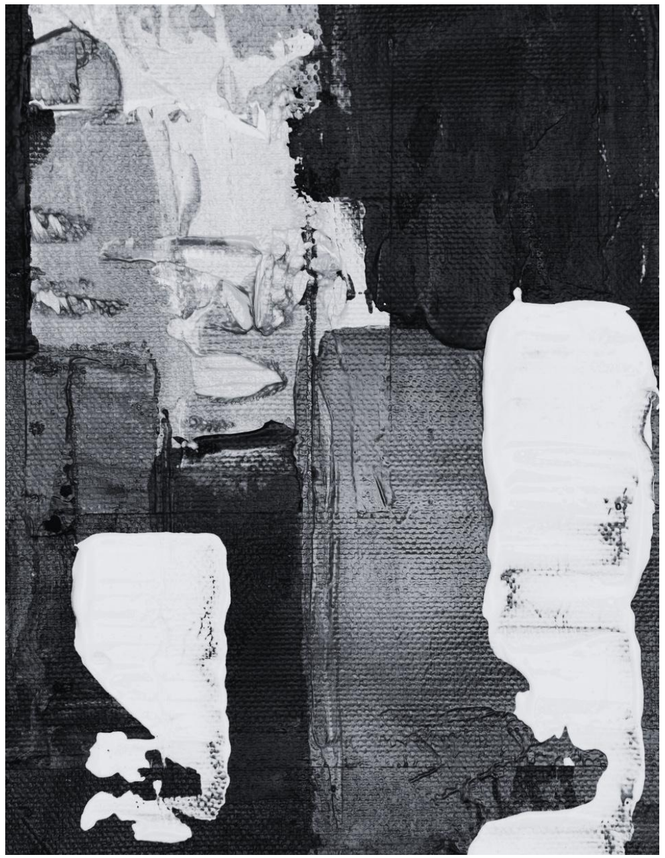
Bien évaluer les usages, les données disponibles et les tendances technologiques

Dans le contexte de l'intégration croissante de l'intelligence artificielle dans les produits, la phase de discovery devient encore plus importante. Les solutions basées sur l'IA nécessitent une compréhension approfondie des données, des capacités de l'algorithme et des implications éthiques de leur utilisation. La discovery aide à définir comment l'IA peut être utilisée de manière efficace, à évaluer les données nécessaires, et à anticiper les impacts de l'intégration de l'IA.

La discovery est d'autant plus cruciale dans l'univers de l'IA où les choix technologiques et stratégiques doivent être réalisés avec une précision et une anticipation des évolutions technologiques importantes dans ce domaine. La notion d’investissement devient alors une question centrale de la stratégie qui se doit d'être éclairée finement.

#Produits #ProcessusProduits #FinOps #GenAI

L’approche FinOps, clef du succès de vos projets GenAI



L'intégration d’une démarche FinOps revêt une importance cruciale dans le parcours des projets GenAI, dès la phase de démarrage, avec le MVP (Minimum valuable product). L’objectif est d'obtenir une estimation précise de la rentabilité potentielle du projet, établissant ainsi des bases solides pour sa transformation en produit pleinement opérationnel.

Les différents projets GenAI sont souvent porteurs de promesses, mais leur concrétisation en tant que produit nécessite une évaluation minutieuse des gains envisagés, tout en anticipant les coûts inhérents à leur développement et mise en production. Pour éviter les déceptions, la clé réside dans le lancement de la phase de MVP avec une version initiale d'un business plan, intégrant la dimension investissement ainsi qu'une estimation pour une mise en production. La prévision des coûts de production amène une réflexion approfondie sur la stratégie de diffusion du produit GenAI, permettant de créer divers scénarios liés au nombre d'utilisateurs et à leur taux d'utilisation.

Grâce à cette première analyse des coûts et à la modélisation des scénarios d'utilisateurs, les équipes obtiennent une vision claire des gains minimaux requis pour chaque cas de figure, permettant ainsi d’arbitrer efficacement sur le projet.

Optimiser la consommation et la facturation

Cette intégration du FinOps dès le début va accélérer significativement la transition vers la phase produit. En maîtrisant dès le départ les coûts et en ayant une vision précise des ambitions du produit, les équipes sont en mesure de franchir efficacement le cap crucial entre le MVP et la phase pleinement opérationnelle. Une planification financière rigoureuse devient ainsi un vecteur majeur du succès des projets GenAI.

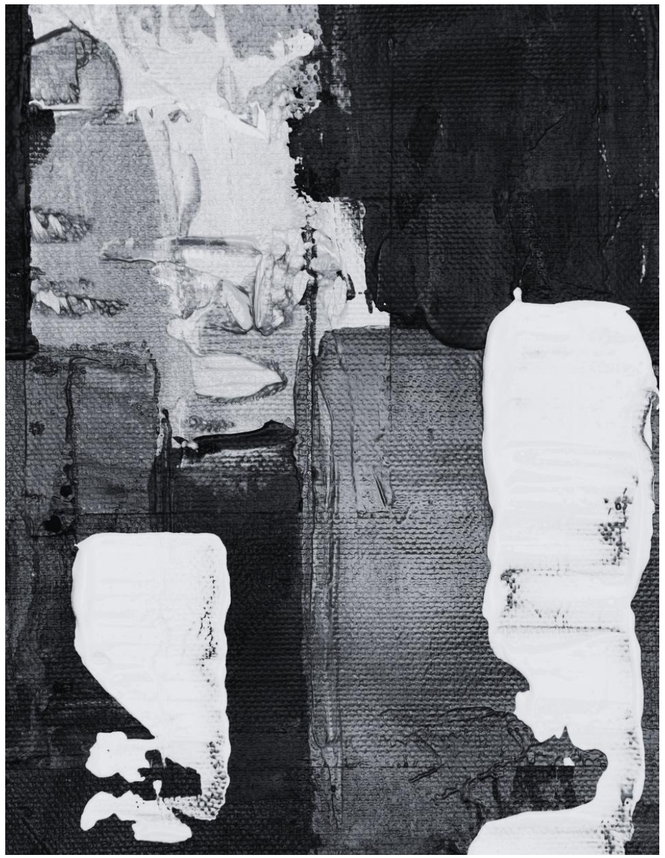
Comme tout projet IT dont la facturation s'effectue proportionnellement à l’usage, les projets genAI bénéficieront d’une approche FinOps basée sur le framework Inform, Optimize, Operate, où les efforts d'optimisation porteront autant sur l’optimisation de la consommation que de la facturation.

PRODUITS #ProductTrends

Ce qui vient dans le monde du produit

#Produits #ProductTrends

De l’approche projets aux modèles opérationnels orientés produits



La valorisation des actifs numériques, de plus en plus présente dans la stratégie de l’entreprise, implique de revoir les modèles opérationnels existants. Les modèles opérationnels orientés produits, qui se focalisent sur la valeur des actifs, sont plébiscités en ce sens. Cette approche marque un changement fondamental par rapport aux modèles traditionnels centrés sur les projets, qui mettent l'accent sur des objectifs et des délais spécifiques.

L’approche produit, pour rester pertinent dans un monde incertain

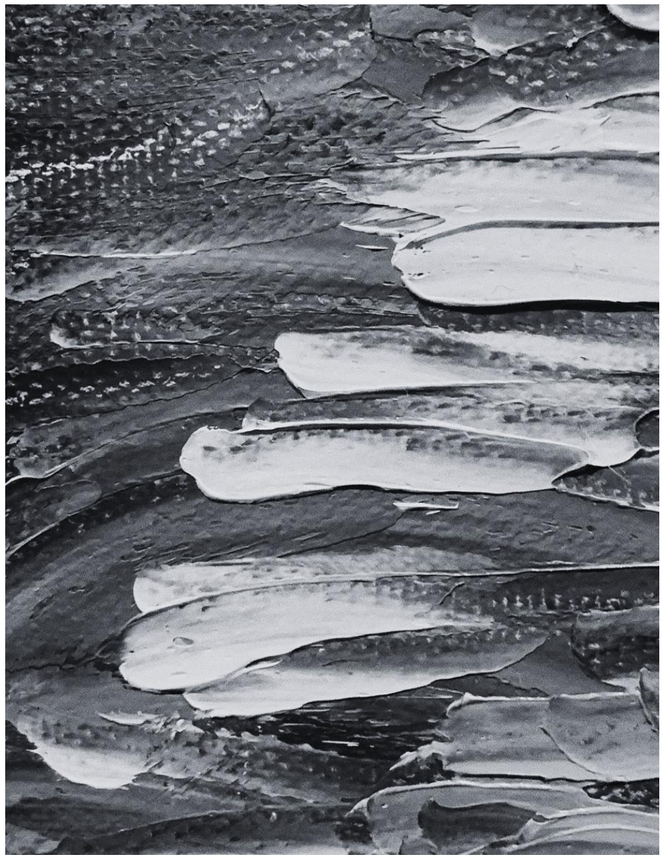
Avec une approche orientée produit, les entreprises peuvent mieux aligner leurs efforts sur les besoins et les attentes changeantes du marché. Elles peuvent ainsi tirer avantages des principes de l’approche pour une plus grande agilité et flexibilité, en s’appuyant sur la responsabilité et l'autonomie des équipes à exécuter la vision et la stratégie de l’entreprise.

D’autres bénéfices peuvent entrer en ligne de compte de l’adoption de cette approche, comme favoriser l’innovation dans un milieu contraint, renforcer la collaboration entre les équipes techniques et les équipes métiers, mieux piloter les actifs et l’investissement, ou encore créer une nouvelle dynamique d’engagement des collaborateurs qui se sentent plus engagés dans leur contribution aux résultats.

L'adoption de modèles opérationnels orientés produit est une réponse stratégique aux défis et aux opportunités qu’ont les entreprises de rester pertinentes, compétitives et en phase avec les évolutions rapides du marché.

#Produits #ProductTrends #PMTech #PMData #PMAi #PMPlatform

Des product managers dans les organisations IT et Data



La prise de conscience des entreprises sur l'importance de valoriser leurs actifs IT & Data comme leviers de différenciation et de croissance engendre des changements organisationnels significatifs. En évoluant vers une approche orientée produit, elles visent à générer plus de valeur à partir de ces actifs. Cette tendance se traduit par l'émergence d'une spécialisation accrue des rôles de product manager adaptée à chaque environnement spécifique et à ses enjeux uniques.

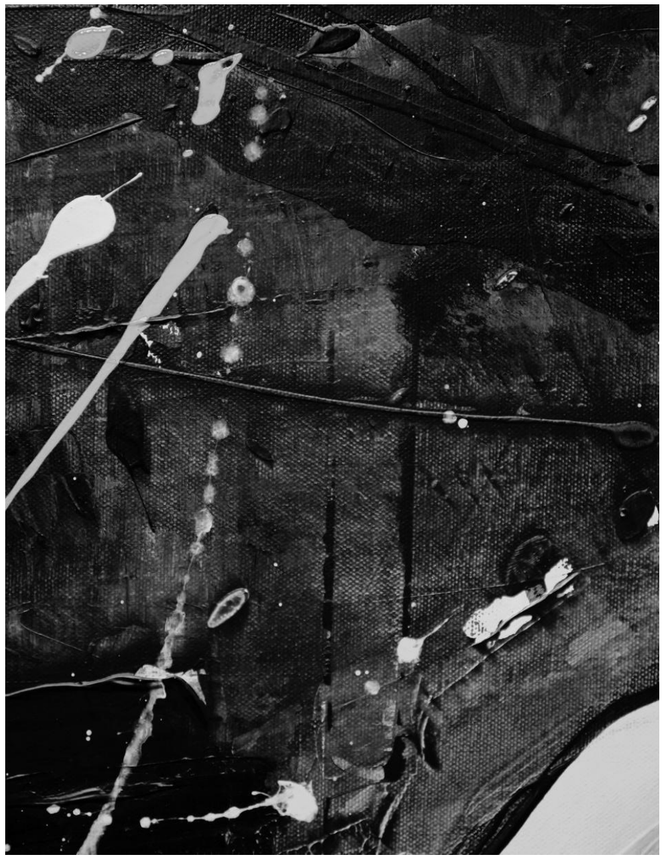
Quatre spécialités se démarquent dans ce contexte : le product manager digital, le product manager data, le product manager IA, et le product manager plateforme.

- Le Product Manager Digital se focalise sur le développement de produits numériques, avec une forte compétence en UX (expérience utilisateur).
- Le Product Manager Data quant à lui valorise les données de l'entreprise pour créer des solutions orientées données favorisant le décisionnel.
- Le Product Manager IA se concentre sur l'intégration de l'intelligence artificielle dans les produits afin de faire émerger de nouveaux cas d'usages.
- Enfin, le Product Manager Plateforme exploite le potentiel des écosystèmes techniques pour en tirer des capacités à forte valeur ajoutée et permettant aux autres équipes de démultiplier leurs capacités.

Cette diversification dans la spécialisation des rôles de product manager reflète l'évolution des stratégies d'entreprise et l'importance croissante d'adapter l'organisation pour tirer pleinement parti de l'introduction de nouvelles technologies.

#Produits #ProductTrends #LeanStartup

Frugalité dans les projets : s’inspirer des concepts clés des start-up



Dans le contexte économique actuel, la frugalité des projets informatiques en entreprise devient une nécessité impérieuse. Les organisations sont de plus en plus confrontées au défi de maîtriser leurs investissements tout en restant compétitives et innovantes. Pour relever ce défi, un changement fondamental dans la manière d'opérer s'impose, et c'est là que les principes du Lean Startup combinés aux pratiques de product management interviennent.

Une approche centrée sur la recherche de l’efficience

Le Lean Startup, avec son approche centrée sur l’efficience (création de valeur maximale avec des ressources minimales), offre un cadre robuste pour le développement de projets informatiques. Cette méthodologie met l'accent sur l'apprentissage rapide, l'itération et la flexibilité. Au lieu d'engager d'importantes sommes dans des projets à long terme basés sur des hypothèses non testées, les entreprises peuvent adopter des cycles de développement plus courts et axés sur les données pour valider leurs idées. Cela minimise les risques financiers et assure une meilleure adéquation entre les produits développés et les besoins réels des utilisateurs.

Parallèlement, le product management offre un autre cadre de frugalité en limitant les ressources (stabilité de l’équipe) et en priorisant l’investissement de ces ressources conformément à la stratégies et aux objectifs de l'entreprise et des attentes du marché. Ce cadre offre également une mesure des résultats, qui favorise la prise de décision fondée sur les données contribuant à l’agilité de l’orientation des projets vers un meilleur ROI.

La combinaison du Lean Startup et du product management offre aux entreprises un moyen puissant de conduire des projets informatiques avec frugalité et efficacité. Cette approche holistique non seulement aide à maîtriser les investissements, mais stimule également l'innovation, la réactivité au marché et, en fin de compte, le succès commercial durable.

#Produits #ProductTrends #GenAI

L’IA générative au centre des questions de la valeur pour l’entreprise

L'année 2023 a été marquée par un engouement sans précédent pour l'IA générative, suscitant un intérêt massif tant dans le grand public que dans les l’entreprises. Toutefois, l'excitation initiale laisse place à des interrogations plus pragmatiques sur les applications réelles de cette technologie et la valeur qu'elle peut apporter aux entreprises.

La fascination des capacités de l'IA générative a occulté la nécessité d’une compréhension claire de son impact réel sur les activités commerciales.

Considérer la GenAI comme un produit pour piloter par la valeur

À présent, les entreprises doivent dépasser cette phase de curiosité pour se concentrer sur un retour sur investissement tangible. Avec des coûts d’utilisation mais aussi d'implémentation pour assurer des défis tels que la sécurisation des données, il devient impératif d'intégrer l'IA générative dans une stratégie de produit bien définie qui saura naviguer entre les enjeux réglementaires, éthiques et techniques tout en exploitant ses opportunités pour l'entreprise.

Cela signifie intégrer l'IA générative non pas comme une fin en soi, mais comme un moyen d'enrichir et de différencier les offres et les services. Que ce soit pour personnaliser l'expérience client, automatiser la création de contenu, ou développer de nouvelles fonctionnalités innovantes, l'IA générative doit être envisagée comme un outil stratégique pour renforcer la proposition de valeur des entreprises.

Les entreprises qui réussiront à intégrer judicieusement cette technologie dans leur stratégie pourront non seulement réaliser un ROI significatif, mais aussi se positionner de manière avantageuse dans un marché en constante évolution.

#Produits #ProductTrends #PMM

Le product marketing manager émerge

Les produits digitaux deviennent de plus en plus des actifs commerciaux de l'entreprise. Pour soutenir la communication entre les équipes Marketing et Tech, un nouveau rôle émerge : celui du Product Marketing Manager (PMM). Ce rôle a une double fonction essentielle.

Premièrement, il fournit les éléments marketing nécessaires pour cadrer la phase de découverte du produit, incluant la segmentation, le positionnement, les études de marché, etc. Ces informations sont cruciales pour définir la direction et la stratégie du produit, assurant ainsi qu'il réponde aux besoins du marché et se distingue de la concurrence.

Un rôle clé pour le succès des futurs produits

Deuxièmement, le PMM travaille en étroite collaboration avec le Product Manager pour élaborer et suivre le plan de lancement du produit. Il s'occupe de la communication et des contenus nécessaires permettant de décrire la valeur et les avantages du produit non seulement aux clients, mais aussi en interne. Il alimente les équipes commerciales avec des informations clés, leur permettant de vendre le produit plus efficacement. De même, il fournit des orientations importantes aux équipes de support, leur permettant de mieux comprendre le produit et d'offrir un service client de meilleure qualité.

Le rôle du PMM est donc au cœur de la stratégie produit. Il agit comme un lien entre le développement technique et les exigences du marché, garantissant que les produits ne sont pas seulement techniquement viables, mais aussi commercialement compétitifs. En assurant une communication fluide et efficace entre les équipes techniques et commerciales, le PMM joue un rôle crucial dans le succès commercial des produits digitaux, les positionnant ainsi comme des atouts stratégiques pour l'entreprise.

#Produits #ProductTrends #DataProducts #AIProducts

Une approche produit aussi pour la Data et l'IA

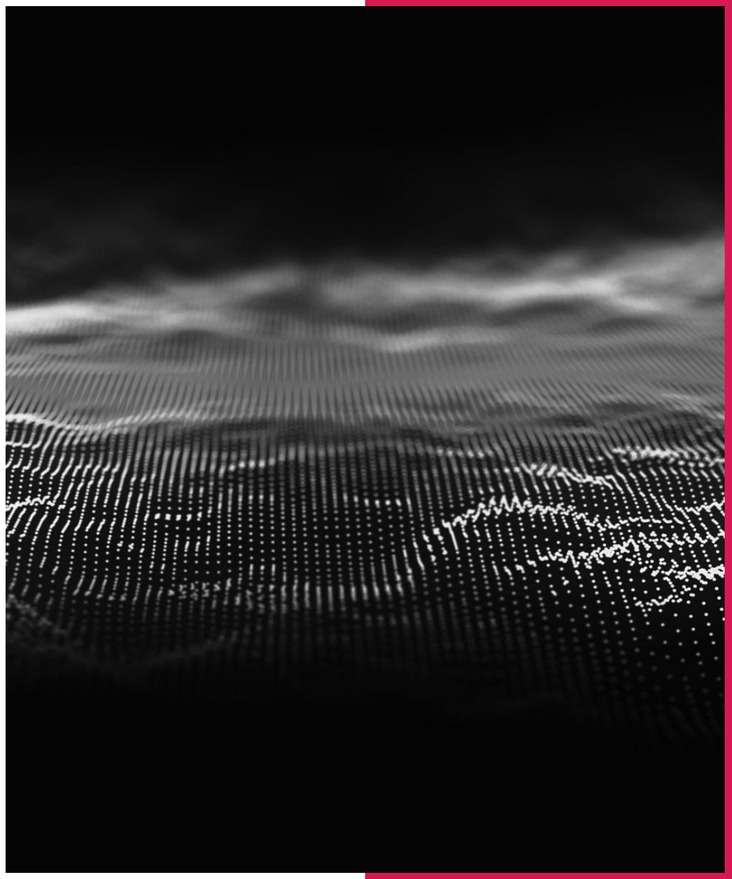
L'approche produit consiste à considérer les données comme un produit qui doit être conçu, développé et commercialisé. Cela implique de mettre l'accent sur les besoins des clients et des utilisateurs, et de travailler en collaboration avec les équipes commerciales pour identifier les opportunités de valorisation des données. Développer une approche produit au sein du pôle Data agit ainsi comme un catalyseur permettant de transformer la donnée en un levier stratégique.

Cette approche produit de la data amène 3 bénéfices clés :

- Une meilleure compréhension des besoins clients. L'approche produit conduit l'équipe data à se concentrer sur les besoins des clients et des utilisateurs, afin de délivrer des produits et des fonctionnalités qui répondent réellement à leurs attentes.
- Une plus grande agilité et réactivité. Dans l'approche produit, l'équipe data est en relation plus étroite avec le métier, et donc mieux à même d'identifier les problèmes métiers sous-jacents, permettant ainsi une réflexion proactive pour fournir des solutions axées sur les besoins concrets du business.
- De nouvelles perspectives pour la valorisation des données. En adoptant une vision centrée sur le client, l'équipe data peut développer des produits et des fonctionnalités qui enrichissent considérablement les services ou produits de l'entreprise,en créant de nouveaux assets stratégiques ou en exploitant de nouvelles niches de marché.

La mise en place d'une approche produit au sein d'une équipe data suppose que la culture soit adaptée : adopter l'orientation produit est une démarche culturelle avant d'être méthodologique. L'équipe data devra soit travailler en étroite collaboration avec une équipe produits, amenant les rôles et les processus adaptés, soit mettre en place ces rôles et processus en interne. Dans tous les cas, des mécanismes de mesure et de suivi devront aussi être mis en œuvre.

DATA



/06

DATA

LES CONTRIBUTEURS DE CETTE SECTION

ILS ET ELLES ONT ÉCRIT LES TECH TRENDS 2024

Ravidhu DISSANAYAKE
Consultant Data / IA,
WENVISION

Martin ELIARD
Head of Data Governance,
WENVISION

Marie FONTAINE
Head of AI,
WENVISION

Florent LEGRAS
CTO Data Groupe,
SFEIR

Mylène MAIGNANT
Consultante IA,
WENVISION

Aurélien PELLETIER
Consulting Director Platform,
WENVISION

Baptiste PUGNAIRE
Consultant Data / IA,
WENVISION

Olivier RAFAL
Consulting Director Strategy,
WENVISION

Céline THOORIS
Consulting Director Data,
WENVISION

DATA #DataConcepts

Les concepts technologiques à maîtriser

#Data #DataConcepts #Architecture

Data Mesh, pour un usage croissant et décentralisé

Le Datamesh est un cadre conceptuel qui englobe plusieurs dimensions et repose sur la notion de fédération. Nous connaissions ses aspects d'architecture du SI : découper les systèmes informatiques en domaines indépendants et autonomes associés à une approche produit, les doter d'une couche d'interface (API ou autre) et de la documentation associée pour que le reste de l'organisation puisse venir consommer les produits.

En 2023, le datamesh s'est particulièrement illustré dans ses impacts organisationnels en entreprise. L'article fondateur de Zahmak Deghani de 2019 propose d'ailleurs avant tout un moyen de sortir de l'impasse aux entreprises qui ont une organisation complexe, morcelée et parfois sillotée. C'est un chemin possible pour ces entreprises, qui n'ont pas réussi à créer un point de vérité central d'entreprise avec une gouvernance forte et un partage exhaustif du patrimoine de données via un datalake maîtrisé et performant.

Le Data Mesh impose une approche Produit, avec une gouvernance appropriée

D'un point de vue organisationnel, le datamesh propose de s'aligner sur la vision Produit, à savoir : chaque unité de production (souvent un domaine data croisé d'un domaine métier) est responsable de bout en bout des produits qu'elle est en capacité de proposer à ses clients. L'unité de production dispose de tous les moyens nécessaires pour produire et est en capacité de prendre des décisions sur les évolutions à implémenter pour servir au mieux ses clients.

La gestion de la plateforme technique reste la prérogative de la DSI, qui elle aussi adopte le mode d'interaction du produit et travaille à mettre à disposition son produit : la plateforme data composée de l'ensemble des éléments de stockage, des flux et outils d'exploitation et surveillance associés au périmètre data.

Deux éléments clés d'un déploiement sans couture d'une organisation data mesh :

- la **data gouvernance** avec des règles et standards uniformes pour l'ensemble de l'entreprise. La data gouvernance peut être traitée par une équipe "chapeau" de l'organisation data si elle a ses points de relais dans les différents domaines. Sinon elle peut être intégrée au dessus de l'équipe plateforme et s'associer aux équipes plateforme.
- la **gestion du patrimoine data** de l'entreprise qui doit être matérialisé dans un set documentaire partagé avec l'ensemble des collaborateurs. La documentation des données, de leurs transformations et des domaines owners est soit décentralisée dans les domaines, soit centralisée au dessus de l'équipe plateforme. La donnée est alors pensée "as-a-service" et constitue une brique complémentaire du mesh.

Le datamesh c'est donc un peu du meilleur des mondes adapté aux enjeux spécifiques de la data pour les entreprises. Le framework est utile et a déjà prouvé son efficacité chez un grand nombre de clients, mais on ne peut faire l'économie d'une gouvernance des données maîtrisée et déployée à l'échelle au risque d'aboutir à des bastions et de retomber dans les problèmes d'enclavement des données.

Le SQL éclipsé par les LLM pour l'interrogation des données

SQL est un langage puissant et polyvalent utilisé pour gérer et manipuler les données stockées dans des bases de données. SQL a pour lui sa simplicité et sa facilité d'utilisation. Il utilise une syntaxe limitée qui permet aux utilisateurs d'écrire rapidement et facilement des requêtes pour extraire, modifier et manipuler les données dans la base de données. Cela en a fait le langage privilégié pour les personnes qui doivent accéder aux données de leur organisation et les analyser.

SQL fournit un langage commun qui peut être utilisé pour accéder aux données et les analyser à partir de diverses sources, notamment les feuilles de calcul, les fichiers plats, les bases de données NoSQL et les entrepôts de données traditionnels. Avec la prolifération des plateformes de données et la quantité croissante de données générées, l'utilisation du SQL est prépondérante chez les Data Analyst, Data Engineer et Data Scientist mais aussi certains profils plus "low-code" comme des business analysts ou des Citizen Developers.

Vers la généralisation des requêtes en langage naturel

Si SQL reste un langage universel, encore largement utilisé et mis à jour, les avancées récentes allant jusqu'à l'intégration de modèles d'IA et de ML directement dans le langage, ces mêmes technologies pourraient progressivement être amenées à le remplacer.

Une des promesses des LLM et de l'IA générative se trouve dans la possibilité pour l'utilisateur d'interagir en langage naturel directement avec un système d'information (le LLM traduisant l'interrogation en requête SQL). Il n'y a donc qu'un pas à franchir pour envisager de questionner directement ses bases de données : "Quels sont les résultats du Q3 ?" ou encore "Liste nos 5 produits les plus importants en terme de CA".

Les géants de la tech comme Google ou Microsoft s'attaquent déjà à ce marché de l'analytique et de la business intelligence, et les premières démonstrations de leurs assistants d'interrogation en langage naturel esquissent déjà les contours d'un futur sans SQL à manipuler - sauf par les experts du data engineering.

#Data #DataConcepts #Lakehouse #DataPlatform

Du Data Lakehouse à la Data Intelligent Platform

Le concept de Data Lakehouse, poussé par des acteurs comme Databricks, Starburst ou Cloudera, vise à apporter “le meilleur des deux mondes” du datalake et du datawarehouse.

Pendant longtemps, les datawarehouses étaient les seuls outils à notre disposition pour stocker et préparer les données structurées aux fins d’analyse de type BI. Les datalakes ont permis d’accueillir beaucoup plus de données, notamment des données non structurées (images, texte, vidéos…). Ils sont aujourd’hui régulièrement utilisés de manière conjointe, au sein d’une Data Platform dans le cloud, le datalake ingérant la donnée brute, dont une partie sera transformée et injectée dans le datawarehouse.

L’idée des Data Lakehouses est de s’appuyer sur un unique produit apportant l’ensemble des fonctionnalités. Toutefois, les opérations de transformation, par exemple, devront toujours exister. Data Lakehouse unique ou Data Platform utilisant conjointement deux produits ou plus, le principe est très proche et devient la pierre angulaire des infrastructures data modernes.

De la GenAI pour intégrer de l’intelligence au sein des lakehouses

Bien sûr, les Lakehouses ayant l’ambition d’agréger des fonctionnalités et de faciliter la vie des utilisateurs, y compris des utilisateurs métiers plus loin du code et de l’informatique, ils regardent aujourd’hui du côté de l’IA générative. Les fournisseurs souhaitent notamment enrichir les capacités suivantes à l’aide de la GenAI :

- **faire du no-code et exprimer en langage naturel les requêtes et fonctions à mettre en place**. L’idée ici est de réutiliser les principes de GenAI adressés aux développeurs via une interface simple d’utilisation. Il ne sera donc plus nécessaire de connaître le langage de la plateforme ou de maîtriser l’ensemble des aspects de data engineering. Le module d’IA générative fera la traduction.
- **des modules de préparation des données intelligents**. Les data scientists n’auraient ainsi plus à passer autant de temps à nettoyer les données, adapter leur format, réaliser des contrôles de qualité. L’IA générative, avec sa capacité à traiter du contenu de la donnée, serait en capacité *a minima* de faire des propositions intelligentes, voire de réaliser les préparations et traitements en amont permettant aux data scientists de se concentrer sur leur algorithme.

Le lakehouse augmenté à l’intelligence artificielle devient donc une Data Platform intelligente et permet à un plus grand nombre d’utilisateurs de s’en saisir, sans limite de compétence techniques.

#Data #DataConcepts #BaseVectorielle

Les bases vectorielles, pour une IA générative d'entreprise

Alors que l'intelligence artificielle générative gagne du terrain dans le monde de l'entreprise, un élément technologique émerge et se distingue par son potentiel disruptif : les bases vectorielles. Véritables piliers d'une intégration sur-mesure des LLM, elles promettent une révolution aussi bien en termes de coût que de performance.

Ces bases ne sont pas nouvelles. Elles sont utilisées pour stocker des vecteurs, c'est-à-dire une représentation mathématique des contenus dans plusieurs dimensions. Elles sont de ce fait très utiles pour du clustering de données, de la recherche par similarité, de la recommandation... Cela en fait des alliées naturelles de l'IA générative en entreprise : grâce à cette technologie, il devient possible de redéfinir la gouvernance et l'exploitation des données en entreprise.

De fait, en entreprise, le véritable intérêt de l'IA générative réside non dans ses propres connaissances, mais dans sa capacité à restituer des informations ancrées dans le vocabulaire et les connaissances de l'entreprise ; on parle alors de "grounding" (ancrage). La principale technique pour cet ancrage est le RAG (retrieval augmented generation), où les documents et données de l'entreprise sont indexés sous forme de vecteurs, au travers du processus d'"embedding". Cela peut se faire à la volée, mais pour un usage du RAG sur des données persistantes, il est commode de stocker ces vecteurs dans une base ad hoc.

Des réponses pertinentes et à jour - si les sources de données sont de qualité

Les entreprises débordent de spécificités qui nécessitent de tels LLM personnalisés. Les bases vectorielles sont donc des alliées privilégiées pour ajouter du contexte dans les interactions afin que les LLM reflètent fidèlement la réalité interne de l'entreprise, sans engendrer les coûts typiquement associés à l'entraînement personnalisé de modèles.

Les "vector databases" promettent donc des analyses finement ajustées aux réalités propres de chaque entreprise. A condition que la qualité des sources soit au rendez-vous, bien sûr ; si l'IA générative s'appuie sur un portail Web qui n'a pas été mis à jour depuis plusieurs mois, par exemple, l'effet risque d'être déceptif.

La question de la sûreté et de la conformité des sources de données est également primordiale. Les bases vectorielles s'avèrent de formidables garde-fous pour garantir et renforcer les politiques de gouvernance de données en contrôlant les accès et en sécurisant les informations traitées de l'entreprise. Elles permettent de plus de garantir l'efficacité des réponses fournies par les LLM (critère lié à la qualité et à la fraîcheur des données). Avec les bases vectorielles, les informations délivrées par l'IA générative peuvent être constamment mises à jour, garantissant des réponses actualisées et fiables - si tant est que les sources soient mises à jour.

#Data #DataConcepts #DataCatalogs

Le data cataloging à l’heure de la GenAI

Avant, les data catalogs n’étaient jamais à jour. Trop fastidieux à remplir. Demain, les métadonnées seront automatiquement mises à jour. C’est l’un des plus beaux cas d’usage de l’IA générative : révolutionner le data management en s’appuyant sur les LLM. Un changement de paradigme aussi important que le passage de l’annuaire Yahoo au moteur de recherche Google.

Le catalogues de données sont au cœur du data management. Un data catalog fournit une vue centralisée et consolidée des données d'une entreprise. Il permet de faciliter la découverte, la compréhension et l'utilisation des données par les collaborateurs de l'entreprise. Plus précisément, un data catalog permet de :

- Décrire les données : source, format, type, définition, qualité, etc. Ces informations permettent aux utilisateurs de comprendre les données et de les utiliser de manière appropriée.
- Répertorier les données : par sources, propriétaires, utilisations, etc. Cela permet aux utilisateurs de trouver les données dont ils ont besoin.
- Rendre les données accessibles en fournissant des outils de recherche et de découverte. Cela facilite l'utilisation des données par les utilisateurs.

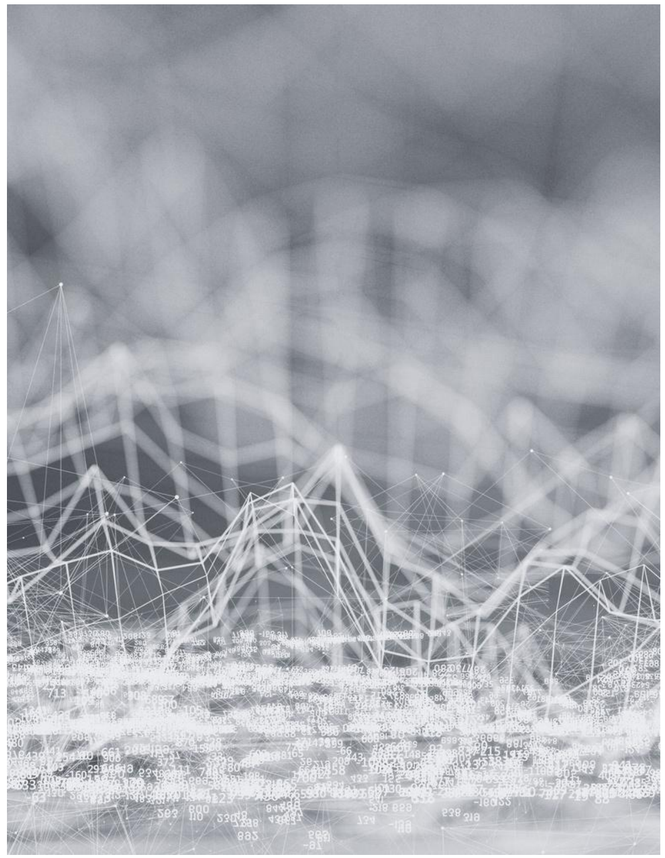
Le problème, c’est que pour rendre ces services, les data catalogs doivent être alimentés et mis à jour manuellement. Un processus laborieux qui ne peut pas suivre le rythme exponentiel de la croissance des données. Ce modèle est donc synonyme d'inefficacité et un fardeau coûteux pour les entreprises, qui induit une certaine défiance vis-à-vis du data cataloging..

Aujourd'hui : l'inspiration Google

L'automation change la donne. Similairement à l'idéologie de Google, un data catalogue piloté par IA mise sur l'automatisation pour crawler, indexer et référencer les données, boostant ainsi la pertinence et la facilité d'accès. Avec leur capacité de traitement du langage naturel, les LLM offrent des solutions pour traduire des volumes massifs de données en informations pertinentes et exploitables, simplifiant considérablement les tâches de recherche et d'analyse pour les utilisateurs à tous les niveaux de l'entreprise.

Face à la complexité croissante et à la quantité de données générées, l’IA générative est non seulement une aubaine mais une nécessité pour réinventer les méthodologies de data management. La GenAI devrait donc devenir le pilier d'un système de navigation des catalogues de données d'entreprise où trouver l'information désirée sera aussi simple que de poser une question à un moteur de recherche grand public.

L’art du Prompt



Le rôle du prompt, ou de l'instruction donnée à un modèle d'IA générative, est devenu un élément-clé de nos interactions avec ces systèmes. Que ce soit à travers les interfaces utilisateur ou par l'écriture directe de prompts par les développeurs, le prompt est la pierre angulaire qui permet de tirer parti de la puissance des LLM pour diverses tâches.

“Prompt Engineering” versus “Prompt ingéniosité”

Le prompt se divise en deux catégories distinctes : les prompts générés par les utilisateurs des interfaces et les prompts écrits par les développeurs. Dans le premier cas, les utilisateurs interagissent avec des modèles pré-entrainés en fournissant des instructions en langage naturel. Dans le second cas, les développeurs interagissent directement avec le code en écrivant des prompts dans un langage informatique. Ce sont eux qui vont itérer sur les prompts afin de trouver les meilleures formulations, qui pourront être mises à disposition des utilisateurs au sein d’une bibliothèque de prompts.

Cela ne veut pas dire que les utilisateurs passant par une interface de chat ne puissent pas faire preuve d’une certaine ingéniosité dans leurs prompts.

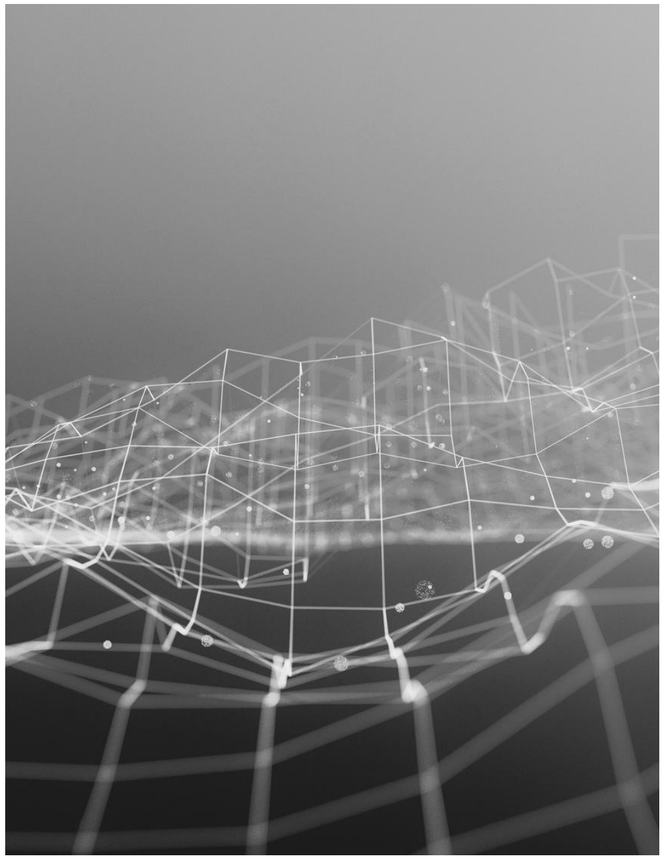
Quelques bonnes pratiques

Le vrai pouvoir du prompt réside dans sa polyvalence. En quelques phrases bien formulées, il est possible d'obtenir un script complet, un résumé d'un document complexe, ou même la génération automatique d'un rapport détaillé. Bien que son utilisation puisse sembler simple, il existe plusieurs bonnes pratiques à suivre pour obtenir des résultats pertinents et efficaces :

- Rédiger des informations claires et structurées
- Donner le plus de détails possibles sur le contexte
- Indiquer le format de sortie désiré
- Expliquer la finalité de la requête

#Data #DataConcepts #GenAIvsML

On n’a pas encore fait le tour du Machine Learning “traditionnel”



Nous avons assisté à l'émergence spectaculaire des Large Language Models (LLM) et de l'IA générative, qui bousculent les perspectives en offrant de nouvelles opportunités : génération de texte, synthèse vocale et créativité assistée. Cependant, elles ne sauraient totalement supplanter les méthodes classiques de ML pour plusieurs raisons.

Premièrement, tous les problèmes ne peuvent être traités avec succès à l'aide de l'IA générative et des LLM. Ces modèles peinent à résoudre des problèmes complexes, à inférer des lois physiques ou à s'adapter rapidement à de nouvelles situations. Deuxièmement, ce ne sont pas des systèmes déterministes : les réponses peuvent varier d’une fois à l’autre. Ensuite, les LLM reposent sur des architectures massives qui sont connues pour leur voracité énergétique et leur empreinte carbone substantielle, ce qui constitue un frein à leur adoption, en particulier dans les segments de marché où la durabilité et l'impact écologique constituent des facteurs de différenciation. Enfin, il est beaucoup plus difficile d'entraîner efficacement un modèle génératif sur ses propres données ce qui limite la personnalisation et l'adaptation à un contexte métier précis.

Le Machine Learning garde toute sa pertinence pour un large éventail de cas d’usage

Compte tenu de ces limitations, les techniques traditionnelles de ML conservent toute leur pertinence et leur valeur ajoutée dans les entreprises, en particulier on peut citer :

- Régression ou classification linéaire : dans les scénarios où les relations entre variables sont relativement simples, stables et linéaires, les méthodes classiques constituent des options privilégiées, en raison de leur facilité d'implémentation, de leur interprétabilité et de leur efficacité.
- Arbres de décision et forêts aléatoires restent des choix populaires dans les applications de diagnostic médical, de scoring de crédit, de segmentation marketing et de maintenance prédictive grâce à leur capacité à représenter des hiérarchies de décisions complexes et leur forte interprétabilité.
- Techniques de clustering, essentielles pour analyser et interpréter de gros volumes de données. Le clustering est fréquemment employé pour identifier et caractériser des segments de clientèle homogènes en fonction de leurs comportements d'achat ou encore pour anticiper les défaillances ou les anomalies en analysant des indicateurs de performance et de santé des machines, installations ou patients concernés.

D'autres applications du deep learning méritent également d'être mentionnées comme la Computer Vision ou le Reinforcement Learning. L'intérêt pour l’IA générative a permis à tous de redécouvrir les possibilités qu'offre le Machine Learning ; c'est l'opportunité pour les entreprises de repenser leur stratégie en y intégrant ces technologies innovantes sans céder aux "buzzword" du moment pour concevoir une stratégie IA cohérente et durable, intégrant ML et GenAI.

DATA #DataGouv

La gouvernance plus que jamais indispensable

#Data #DataGouv #DRHData

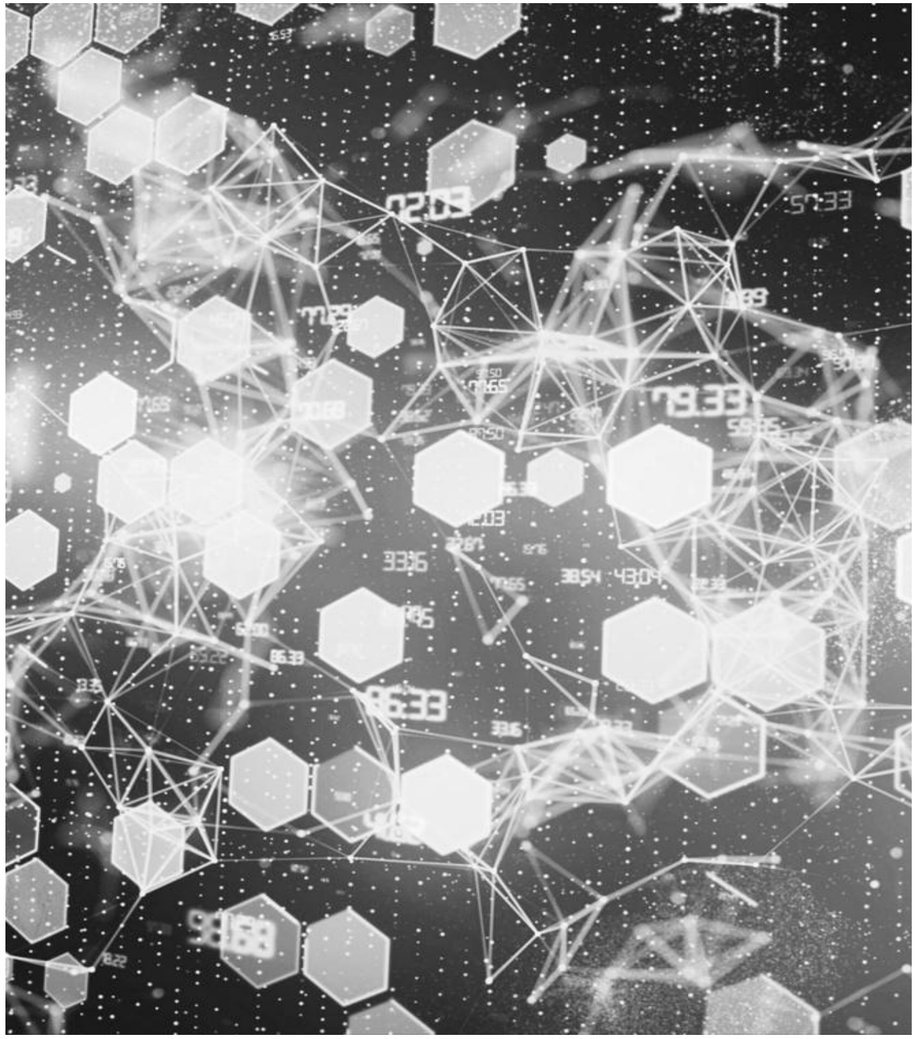
Des DRH Data pour catalyser les organisations Data

Dans le paysage data-driven de 2024, un rôle majeur doit prendre enfin toute sa place : celui du Directeur des Ressources Humaines Data. Son implication devient incontournable pour piloter la stratégie talent au cœur des organisations. En collaboration avec le CDO et les équipes Data, il est le catalyseur essentiel pour fédérer les compétences et acculturer les collaborateurs à la valeur stratégique des données, un atout indispensable pour les entreprises qui souhaitent pérenniser leur organisation.

Les rôles essentiels d'un ou une DRH Data

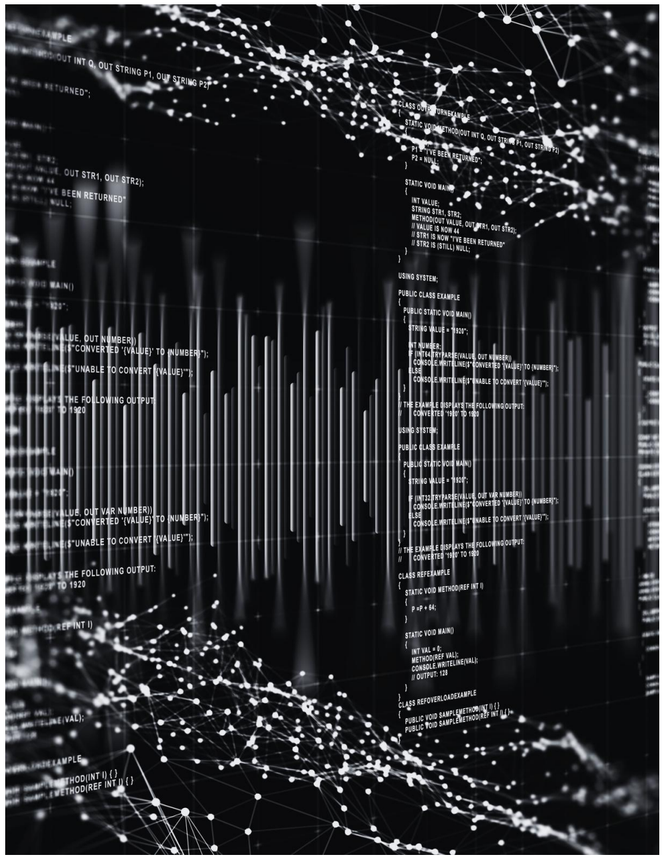
- **L'intégration des talents.** Le DRH Data endosse le rôle crucial d'intégrer les talents essentiels aux besoins de l'organisation. Sa mission commence par une cartographie détaillée des postes et responsabilités, fondamentale pour le déploiement d'une gouvernance des données pérenne.
- **La formation des talents.** Conscient de l'évolution constante des technologies, il met en œuvre des programmes de formation personnalisés. Il s'assure que chaque membre de l'organisation, quels que soient son rôle et son niveau, accède aux formations nécessaires pour s'adapter et exceller dans son domaine. Ces initiatives soutiennent et stimulent le déploiement d'une culture data globale.
- **La reconnaissance des talents.** Le DRH Data veille à instaurer une culture du feedback où la reconnaissance est centrale. Signe distinctif d'une culture d'entreprise progressiste, il valorise les réalisations individuelles et collectives et offre des possibilités d'évolution à travers un parcours professionnel riche et data-orienté. La reconnaissance motive les talents et stimule un sentiment d'appartenance à l'entreprise.

Investir dans le rôle du DRH Data est une étape décisive pour toute entreprise ambitieuse. C'est lui qui va piloter l'asset le plus précieux des organisation data driven : le capital humain.



#Data #DataGouv #GenAI #Rôles #ML

Quels rôles pour inclure l'IA générative dans sa stratégie Data ?



L'essor de l'IA générative et des LLM renforce le besoin de déterminer précisément les rôles et responsabilités en matière de data et d'IA, ne serait-ce que pour anticiper les besoins en recrutement et formation. Quelques tendances notables sont à surveiller comme l'augmentation de la demande en expertise en IA générative et en LLM, l'émergence de profils de plus en plus hybrides gardant un lien fort avec les équipes métiers mais bénéficiant de nouvelles capacités augmentées par l'IA, le renforcement de l'importance de l'éthique et de la transparence. Les professionnels de la data et de l'IA seront tenus de redoubler d'efforts pour garantir l'honnêteté, l'impartialité et la traçabilité de leurs modèles.

Voici un aperçu des principales fonctions et attributions associées à chacun d'entre eux :

- Les **Data Analysts** manipulent la donnée pour en tirer des enseignements clés, afin de résoudre des problèmes du business ou d'aider à prendre des décisions mieux informées. Ce rôle est amené à s'éloigner de la DSI pour être plus proche des métiers car les LLM contribuent à faire sauter la barrière technique.
- Les **Data Scientists** construisent des modèles mathématiques de machine learning pour répondre à des problématiques métiers. Dans la majorité des cas, ils s'appuient sur des modèles existants qu'ils personnalisent pour répondre à des enjeux opérationnels, il aussi seront en charge de la veille technologique en l'absence de profils dédiés à la recherche.
- Les **Machine Learning Engineers** se concentrent sur le déploiement et l'opération des modèles d'apprentissage automatique dans des environnements de production. Leur principal objectif est de traduire les modèles développés en solutions logicielles robustes et évolutives. Il doivent être moteur pour la mise en place de la démarche MLOps et bientôt de LLMOps afin de détecter automatiquement toute dérive.
- Les **Data Engineers** développent et maintiennent les outils et infrastructures permettant l'analyse de la donnée. Ce sont eux qui vont implémenter les pipelines de données et permettent sa circulation efficace au sein d'une entreprise.
- Les **AI Research Scientist**, qu'on retrouve chez les entreprises les plus innovantes et ambitieuses, se consacrent à la recherche fondamentale en IA, explorant de nouvelles idées, concepts et approches pour résoudre des problèmes complexes. Leur travail doit faire rayonner l'excellence technique de l'entreprise tant via des prises de parole dans des conférences prestigieuses que par des publications de haut niveau.

#Data #DataGouv #Gouvernance #GenAI

La gouvernance de la donnée, gardienne de la fiabilité des LLM

La montée en puissance de l'IA générative, illustrée par les avancées de modèles comme Gemini ou GPT-4, a marqué un tournant technologique majeur, prompt à transformer notre façon d'interagir avec les données. À l'horizon 2024, la gouvernance de l'IA s'impose comme la pierre angulaire permettant de tirer pleinement parti de ces technologies de pointe, tout en atténuant leurs risques inhérents.

La gouvernance de l'IA est un investissement capital pour les entreprises souhaitant naviguer avec succès dans l'écosystème de l'IA en 2024. Elle est la garante d'une technologie IA générative sécurisée, conforme et éthique, assurant tant l'excellence opérationnelle que la valorisation des données.

A contrario, de nombreux cas concrets initiés en 2023 démontrent comment une gouvernance de données défaillante peut mener à des incohérences opérationnelles, telles que des recommandations de produits erronées, ou encore des accès à une donnée non autorisée. Ce constat souligne l'impératif d'une gouvernance adéquate pour l'intégration réussie de l'IA générative dans les processus métier.

Les grands principes à garder en tête

- **La qualité des données comme socle de la fiabilité**

La réussite des systèmes d'IA générative repose sur un adage intemporel : une qualité de données irréprochable conditionne l'exactitude des résultats. Une gouvernance de données robuste assure l'intégrité et la sécurité des datasets, garantissant ainsi fiabilité et pertinence des LLM.

- **L'éthique pour une utilisation sécurisée et responsable**

La gouvernance de l'IA doit veiller à la conformité des données vis-à-vis des réglementations en vigueur, notamment le RGPD en Europe, et s'assurer du respect de la vie privée. Elle doit également prévenir les biais et veiller à l'équité, amplifiant le caractère éthique de l'utilisation de l'IA.

- **Traçabilité et transparence comme leviers**

En instaurant des politiques transparentes et une traçabilité des décisions, la gouvernance de l'IA favorise une responsabilité claire et répond aux exigences réglementaires. Un tel cadre renforce la confiance des utilisateurs et des parties prenantes.

#Data #DataGouv #Stratégie

Les stratégies data et GenAI à l'épreuve de la valeur

Développer une stratégie data de nos jours est essentiel pour exploiter la valeur des données "dormantes" au sein d'une entreprise, d'autant plus que la démocratisation de la GenAI transforme les cas d'utilisation et l'accès aux données.

Pas de valorisation sans plateforme data

Il est crucial de souligner que l'utilisation des algorithmes GenAI, en dehors de l'expérimentation, requiert une plateforme data. À l'instar de la construction d'une maison, les cas d'utilisation pour valoriser la donnée doivent reposer sur des fondations solides, représentées par une plateforme data.

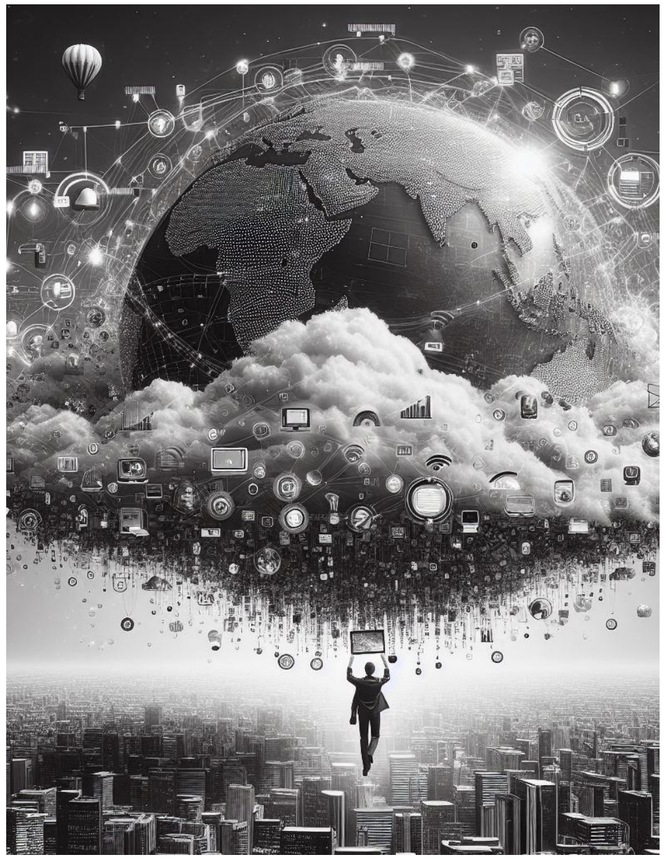
La valeur d'une plateforme : la clé de la pérennité

Dans cette ère de changements technologiques et de conditions économiques incertaines, l'évaluation claire de la valeur d'une plateforme data est primordiale. Les entreprises doivent éviter les écueils classiques liés aux nouvelles technologies :

- ne pas se focaliser uniquement sur l'aspect technique de la GenAI ;
- ne pas en faire un miroir aux alouettes, avec des promesses de millions économisés ou produits en plus ;
- ne pas oublier que la qualité de la donnée en entrée sera déterminante pour le résultat final.

Il est impératif de revenir à la pertinence des cas d'utilisation et de piloter la stratégie data en se concentrant sur la création de valeur. Cette approche revêt une importance particulière, car les algorithmes GenAI, bien qu'ouvrant de nouvelles perspectives, sont gourmands en ressources.

Les concepteurs de plateformes data doivent rationaliser l'utilisation des technologies et définir des KPI clairs quant au coût et à la valeur créée par la plateforme. Ils doivent également intégrer toute la chaîne amont (préparation de la donnée, maintien des bases de connaissance...) dans leurs calculs. Dans une industrie aussi dynamique que la data, seul un pilotage axé sur la valeur permet d'assurer la stabilité et la durabilité d'une plateforme.



#Data #DataGouv #GenAI

L’IA Générative pour améliorer la qualité de la donnée

Actuellement, les algorithmes génératifs sont en passe de révolutionner la philosophie des plateformes de données. Les applications dans le domaine de la data et de l’IA traditionnelle sont nombreuses.

Un exemple significatif est la transformation de la Business Intelligence (BI) grâce à la capacité de ces algorithmes à comprendre les questions liées aux activités business et à extraire les données pertinentes pour y répondre, redéfinissant ainsi les perspectives de la self-BI. La capacité d'interagir avec des données structurées et non structurées, via des algorithmes multimodaux ou la combinaison de modèles, permet une vision simplifiée des données du SI. Toutefois, il ne faut pas se leurrer : le passage à l’échelle de tels cas d’usage est complexe, en grande partie du fait du travail que cela demande sur la préparation de la donnée.

La génération de la donnée au cœur d’un plateforme plus fiable

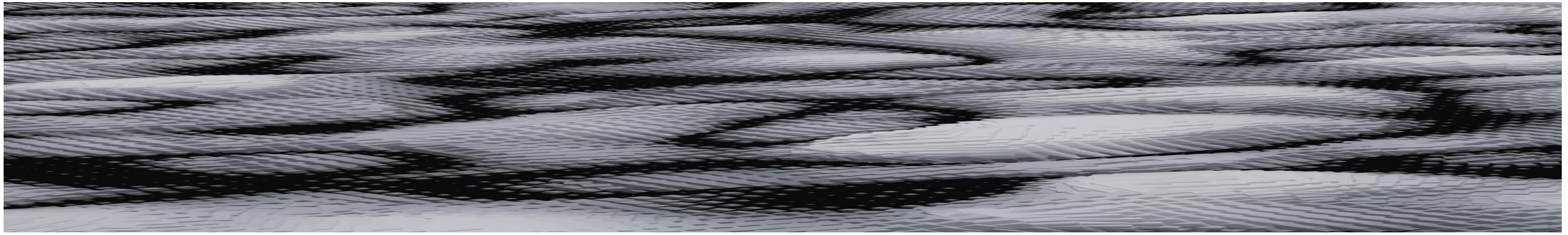
Il faut garder en tête cependant que l’IA générative peut elle-même contribuer à la qualité de la donnée. Les algorithmes génératifs sont capables de déceler des données manquantes mais aussi de créer de nouvelles données à partir de données existantes, ce qui ouvre de nombreuses possibilités dans le domaine de la data. Les données synthétiques peuvent ainsi être utilisées pour générer des jeux de tests pour fiabiliser les pipelines de données afin de s'assurer qu'ils fonctionnent correctement avant d'être déployés en production.

La création de données offre des opportunités prometteuses pour améliorer la gestion des données personnelles. Les développeurs peuvent travailler avec des données générées pour préserver la confidentialité des informations. De plus, la recherche et le tagging automatiques des données permettent un suivi plus efficace et renforcent la sécurité des plateformes de données. En ce qui concerne le développement, la capacité de générer des tests automatiques et une documentation contribue une fois de plus à renforcer la fiabilité des développements.

La donnée générée peut aussi servir de base d’entraînement pour des algorithmes de classification. De nombreux cas d’usage ont émergé autour de la génération d’images, pour entraîner des IA à distinguer des multitudes de formes, états ou variations d’un objet ou d’une situation.

#Data #IA #AgentsAutonomes

L’automatisation à l’ère des LLM : les agents autonomes



Dans le domaine en constante évolution de l'intelligence artificielle, les modèles LLM tels que GPT-4 sont en train de transformer radicalement la notion d'automatisation. Leur impact s'étend bien au-delà des tâches simples, touchant également des processus complexes. **Les modèles LLM donnent une nouvelle dynamique aux agents autonomes en leur octroyant la capacité de segmenter une vaste gamme de problèmes en sous-problèmes plus gérables.** Cette compétence essentielle leur permet de planifier méthodiquement et de construire une suite d'action menant à une solution.

Deux études notables illustrent la progression dans ce domaine. La première, Toolformer, montre que les modèles de langage peuvent apprendre de manière autonome à utiliser des outils externes pour améliorer leurs performances dans une variété de tâches. La seconde, Voyager, présente un agent autonome dans Minecraft, capable d'apprendre et d'évoluer de manière continue sans intervention humaine, en utilisant GPT-4 pour générer et améliorer son propre code et ses compétences.

Toutefois, l'intégration de ces technologies avancées soulève des défis significatifs, notamment en matière de sécurité des données et de fiabilité des systèmes. La gestion des informations sensibles par ces agents autonomes exige des mesures de sécurité robustes pour prévenir les abus et les violations de données.

En entreprise, une approche itérative et de l’accompagnement

Dans le contexte professionnel, l'application de ces technologies LLM transforme la manière dont les entreprises opèrent. Ces outils simplifient le traitement de grands volumes de données, facilitant ainsi l'acquisition de connaissances pour les décisions stratégiques. De plus, la capacité de communication en langage naturel des agents autonomes ouvre de nouvelles possibilités pour l'amélioration de l'expérience client et l'efficacité des services. A noter toutefois que cette automatisation accrue amènera la nécessité de former les employés pour qu'ils puissent tirer pleinement parti de ces nouvelles technologies.

La complémentarité des agents et des LLM inaugure une ère d'automatisation et d'innovation. Comprendre et intégrer ces technologies de manière stratégique est crucial pour maximiser leur potentiel au sein des entreprises, tout en abordant de manière responsable les implications éthiques et pratiques. Cette transition vers une automatisation plus intelligente nécessite une approche itérative et bien informée, pour assurer une adoption bénéfique et durable de ces avancées technologiques.

Remerciements

à l'ensemble des contributeurs de ces TechTrends, merci à vous de nourrir ces réflexions, au quotidien,

à tous nos lecteurs, merci pour vos feedbacks et vos encouragements,

à tous les clients de SFEIR et WENVISION, qui nous font confiance pour guider leur stratégie et mettre en œuvre nos préconisations !

#TheEnd

TENDANCES TECH 2024

Un grand merci de nous avoir lu.
Vous êtes d'accord ? Vous aimeriez discuter ou approfondir un point ? Contactez-nous, venez en parler avec nous, nous serons très heureux de poursuivre l'échange !
Et nouveauté cette année, nous mettrons régulièrement à jour ce document, afin qu'il reste un document de référence tout au long de l'année.

contact@sfeir.com
contact@wenvision.com

[sfeir]
WENVISION