

HOMMAGE À LOUIS NAUGÈS

Cet article a été co-écrit avec Louis Naugès, l'inventeur de la bureautique, 50 ans d'inventions, d'écrits, de réflexions, dernièrement Chief Strategy Officer de Wizy.io pour aider les « frontline workers » (FLW). Malheureusement en ce mois de Mai 2024, Louis nous a quittés. Cette parution lui est dédiée pour lui rendre hommage. Nous espérons que la synthèse et les illustrations seront à la hauteur de ses attentes, à présent qu'il nous regarde depuis le « Cloud ».



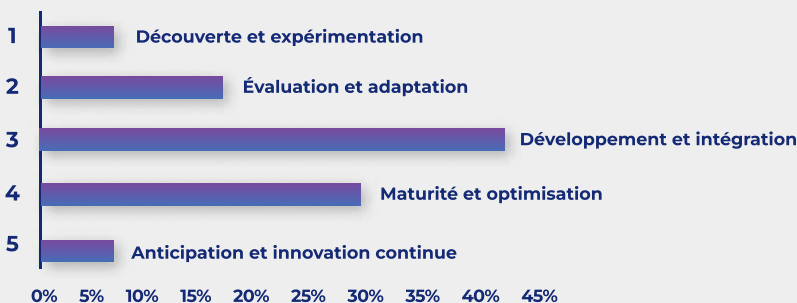
Credit : Paola Naugès - Le Faucheur 



Au-delà du buzz suscité par l'IA Générative, les cas d'usage IA sont en ébullition dans toutes les entreprises, administrations, pôles de recherches et universités.

Le rapport « **AI for Green & Green AI** » rédigé par Ana Semedo, responsable IA de l'Institut G9+* donne un bon aperçu du niveau d'adoption et de maturité de l'IA au sein d'un top 45 de grandes organisations :

Adoption de l'IA (pourcentage d'entreprises, niveau d'adoption 1 à 5)



Qu'en est-il vraiment du rapport Bénéfices/Impacts sous l'angle du FinOps et du GreenOps ?

Comment concilier ces 2 objectifs antagonistes
sans tomber dans les mêmes travers que dans le passé?

Risque-t-on, sous prétexte du « Tech for Good », de consommer inutilement des ressources (coûts, énergie, CO2) pour des **cas d'usage vraiment challengeables dans leur utilité ?**

Dans cette newsletter, on va tâcher de comprendre comment concilier IA, maîtrise des coûts et sobriété numérique en 3 étapes :

- 1 Un rappel FinOps
- 2 FinOps for AI
- 3 AI for FinOps

1 COMMENT MESURE-T-ON le FinOps ?

Pour rappel, le FinOps est un cadre opérationnel et une pratique culturelle qui maximise la valeur Métiers du Cloud en permettant une prise de décision rapide basée sur les données et créant une responsabilité financière grâce à la collaboration entre les équipes d'Ingénieurs, la Finance, et les Métiers.



Quelle est la photographie de l'**Optimisation des coûts et usages du Cloud avant l'IA** ? Au travers des phases « **Inform**er », « **Optimiser** », « **Opérer** », les entreprises cherchent à obtenir des gains tangibles et mesurables via des KPIs, parmi lesquels :

Être en mesure d'allouer à des Centres de Coûts 50% à 90%+ des dépenses Cloud

Bénéficier de remises (Discounts) couvrant entre 60% à 80% de leurs ressources

Avoir la capacité de maîtriser les écarts entre les dépenses prévisionnelles (Forecast spend) et les dépenses réelles dans une fourchette allant de 20% à 12% pour les plus matures.

3 NIVEAUX DE MATURITÉ



Crawl

Walk

Run

Et les critères associés :

- **Fiabilité du Reporting** et de l'Outillage
- **Données** permettant des prises de décision opérationnelles
- **KPIs** pour mesurer le succès
- **Procédures et processus** établis et communiqués
- Compréhension et adhésion aux principes FinOps

Or l'IA change la donne.

Et comme l'illustre Louis dans son exposé, on s'attend à des usages indirects majoritaires de l'IA via des usages indirects sur les plateformes propriétaires des fournisseurs.

ENTREPRISES : 2024 - 2029

Usages directs
15%

OpenAI Gemini

Usages indirects SaaS
75%

Microsoft 365 Copilot
Google Workspace
workday.
HubSpot

Usages coeurs métiers
10%

MISTRAL AI



2

FinOps for AI

COMMENT ANTICIPER LES NOUVEAUX COÛTS LIÉS À L'IA ?

Actuellement, les entreprises déployant des offres IA s'appuient sur 3 principaux modèles de déploiement combinant un compromis entre qualité, confidentialité, précision, performance et coûts :

1

Services à Codes propriétaires de fournisseurs tiers

2

Services à Codes Open Source hébergés par des fournisseurs tiers

3

« Dév Spécifiques Maison » basés sur des Services/Systèmes Centrés IA des Cloud Providers (Do It Yourself : DIY)

Et lorsqu'on entend fournisseurs tiers (Third Party Vendors), c'est là qu'on mesure la position ultradominante des Big 3 : Amazon AWS, Microsoft Azure, Google GCP...

Parmi les exemples les plus populaires de chaque catégorie, on peut citer :

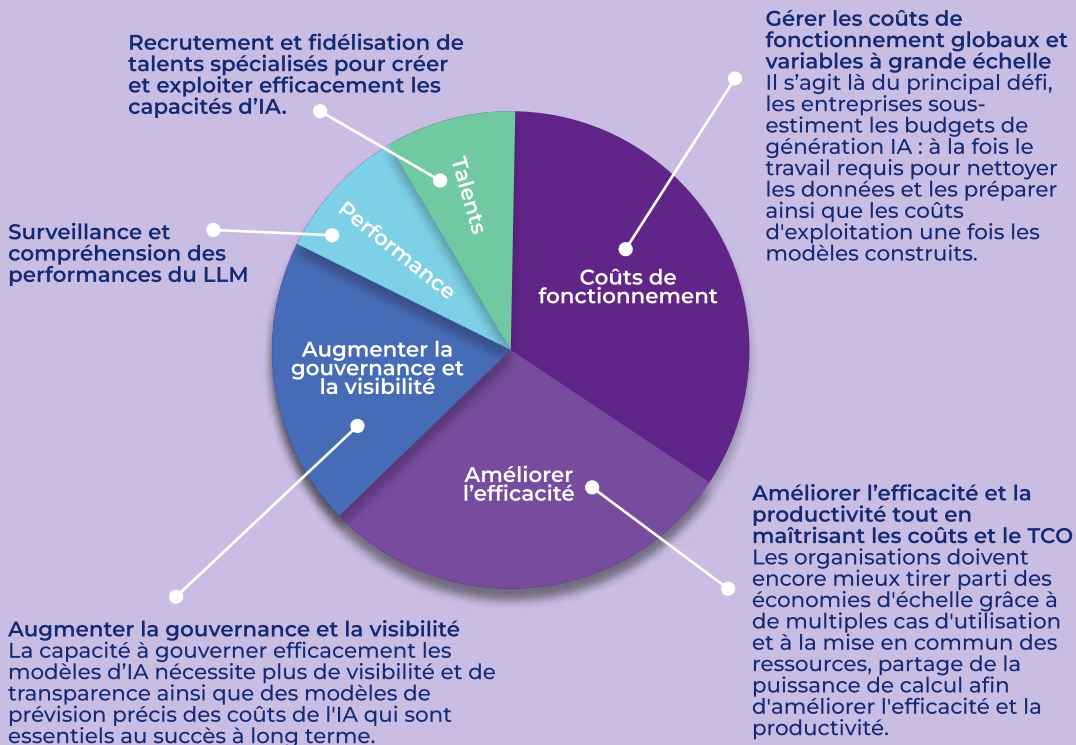


De quoi sont constitués les inducteurs de coûts de ces modèles ?

	Codes propriétaires	Codes Open Source	Do It Yourself (DIY)
Métriques	"Pay-per-Token Model": "tokens in/out" (ex. mots d'un texte) + éventuel Set-Up/ Training	Les coûts sont basés sur le modèle d'entraînement	Coût matériels à ne pas sous-estimer*
Capacité dédiée	Pendant la période de commitment	Volumes processés et stockés	Contrôle complet de la donnée
Propriété Intellectuelle (IP)	Limitée pour le client	Faible	Propriété complète
Compétences et Expertise techniques	Faibles	+ Avancées pour la personnalisation	Expertise avancée
Codes Produits (SKUs)	Peu: la plupart en PaaS	Modèles, données, appels API	Hardware, software modeling, training, APIs et les licences en mode « Self-Hosting Model »
Prix	Contractuels complets sur les usages	Contractuels accès infrastructure et facturation des ressources de calcul (Compute) à la charge du client	Infra & Compute à la charge du client

**Exemple cité par Finout: En DIY (Self-Hosting Model) la principale dépense est le matériel. Par exemple, le déploiement d'un modèle open source tel que Falcon 180B sur une plate-forme comme AWS impliquerait généralement l'utilisation d'un type d'instance tel que ml.p4de.24xlarge, qui peut coûter environ 33 \$ par heure pour une utilisation à la demande. Cela se traduit par une dépense mensuelle minimale d'environ 23 000 \$, hors ajustements de type Auto-Scaling et Discounts.*

Quels sont les premiers défis FinOps de l'IA constatés sur le terrain ?



Comprendre ces défis clés à l'avance et prendre des mesures proactives pour atténuer efficacement les problèmes de coûts est **essentiel au succès à long terme de la Gen AI** au sein d'une organisation.

En fin de compte, on en revient aux fondamentaux du **FinOps**, pouvant être poussés à l'extrême dans le contexte IA :

Provisioning

Les tâches d'IA altèrent l'approche traditionnelle du **Provisioning de ressources** et du **Right-Sizing** : des ajustements plus réguliers que dans le monde traditionnel sont nécessaires.

Instance

La **typologie correcte d'Instances** est beaucoup dépendante des périodes d'utilisation et des pics d'usage : l'utilisation **d'instances Spot est essentielle**.

Scalabilité

Scalabilité : de par la nature fluctuante des workloads d'IA, mettre en place des dynamiques d'Auto-Scaling est parfaitement bien adapté.

Monitoring

Ajustements de Monitoring : Plus que jamais, une surveillance de l'infrastructure dédiée à l'IA est indispensable pour permettre notamment une correcte allocation des coûts (notamment via le Tagging).

Stockage des données

De par la création de volumes de données par essence significatifs avec l'IA, les coûts de stockage peuvent exploser, d'où l'idée de contrôler et maîtriser les patterns avant/ après chaque nouveau projet d'IA.

Transfert

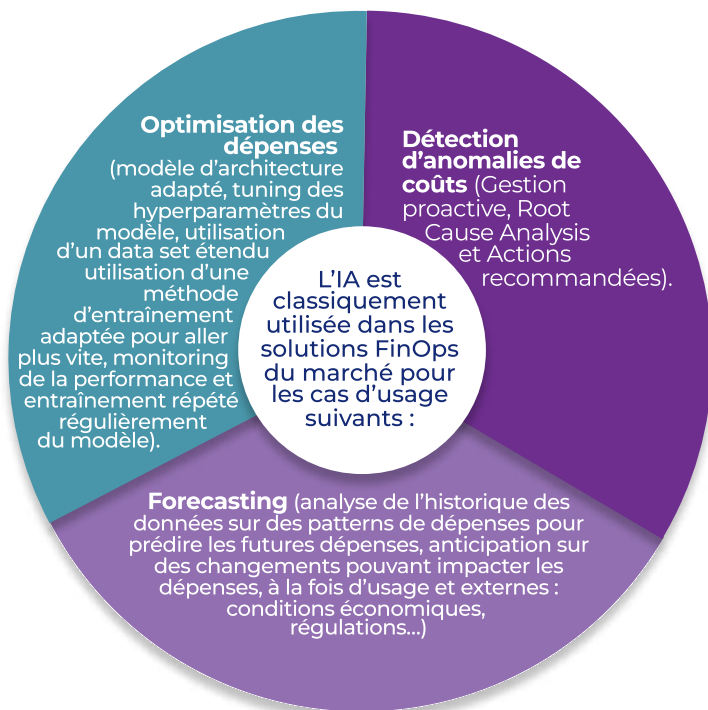
Là aussi, la dimension internationale de certains modèles d'IA peut induire des coûts entre régions Cloud. Utiliser le « Content Delivery Network » des Cloud Providers et l'automatisation des transferts de données entre plateformes (en cas de multi-Cloud) se révèle essentiel.

3

AI for FinOps

UTILISER L'IA POUR UN PRÉVISIONNEL FIABLE ?

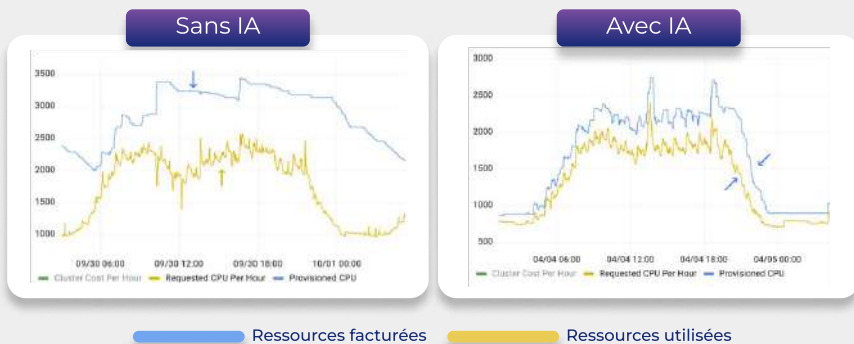
A présent qu'on a abordé les principaux grands modèles, les coûts associés et les moyens de les contrôler du mieux possible, intéressons-nous au dernier sujet du moment dans le monde FinOps : l'utilisation de l'IA pour les modèles FinOps.



L'IA au secours de vos budgets pour le Cloud.

L'IA pour libérer du budget ... pour l'IA.

Comme les budgets des entreprises ne sont pas extensibles et les fournisseurs Cloud l'ont bien compris, il existe en ce moment un encouragement de ces derniers à utiliser des modèles d'IA pour le FinOps avancés notamment dans l'exemple ci-dessous pour le Right-Sizing et l'Auto-Scaling de clusters Kubernetes, afin de libérer 30, 40 voire 50% des coûts facturés pour des ressources sous-utilisées :



Exemples fournis à partir de la plateforme CAST.AI

Citons les classiques IBM Apptio Cloudability, Flexera, CloudHealth qui sont les 3 technologies les plus répandues et qui ont les moyens R&D de pousser loin les modèles d'IA.

A leurs côtés des cloud natives innovantes sortent du lot : citons **Finout et CAST.AI** qui l'une et l'autre ont une approche très avancée de l'IA pour le FinOps.

 **En conclusion,** plus que jamais avec l'IA, l'optimisation des performances, des coûts et de l'empreinte carbone **sont un sujet stratégique dès maintenant pour les entreprises.** Celles qui adopteront d'emblée les meilleures pratiques FinOps / GreenOps dès l'initialisation de leurs grands modèles seront mieux parées d'être à l'abri de mauvaises surprises et auront embarqué leurs collaborateurs dans **une approche saine.**

 **Et si on laissait le mot de la fin à Louis,** il n'aurait pas manqué de nous rappeler sa conviction du Tsunami sur les emplois informationnels mais que **l'IA va aussi aider à revaloriser les Travailleurs Terrain (FLW),** qu'on en est au stade 1 ANI (Narrow) qui précède AGI (General) et ASI (Super), qu'il faut remplacer la peur de perte de **« Souveraineté Numérique »** par l'aspiration à bâtir une **« Confiance Numérique »** et qu'enfin tout reste à faire et que l'Europe avec ses Talents a un rôle majeur à jouer dans l'industrie du Numérique à l'heure de l'IA, à l'image de l'Estonie qu'il aimait prendre en exemple.

 **Cloud FinOps** est un cabinet **100% dédié FinOps indépendant** intervenant en France et à l'international qui aide ses clients à « **booster la Valeur du Cloud** » et à vraiment « **contrôler leurs dépenses Cloud** ».



Cloud FinOps



cloud-finops.io



Télécharger notre livre blanc

Sources

- 1 - CIGREF ([Lien](#))
- 2 - Google Cloud AI for FinOps ([Lien](#))
- 3 - FinOps Foundation FinOps for AI ([Lien](#))
- 4 - Finout FinOps for GenAI ([Lien](#))
- 5 - CAST.AI AI for AI ([Lien](#))
- 6 - Conférence de Louis Naugès à Grenoble École de Management ([Lien](#))

[Se désinscrire](#)