Cloud souverain : l'Intelligence Artificielle au coeur des enjeux

28 NOVEMBRE 2022

Auteur

LAURENT DAUDET

Professeur à l'Université Paris Cité, Directeur général de LightOn

Laurent Daudet a été co-coordinateur du groupe de travail Enseignement Supérieur et Recherche de Terra Nova, entre 2012 et 2017.

Entrepreneur dans le domaine de l'Intelligence Artificielle, l'auteur réagit ici à une précédente note publiée dans La Grande Conversation par Thomas Reynaud ("Pour une souveraineté européenne sur le Cloud et les données", 15 novembre 2022). Il montre l'ampleur des développements autour de l'IA et fait part des besoins industriels liés au Cloud pour réussir à développer des projets de pointe dans ce secteur.

Directeur général d'une startup de très haute technologie dans le monde de l'Intelligence Artificielle (IA), dont le déploiement commercial est principalement porté par les technologies du Cloud, j'ai lu avec grand intérêt la note de Thomas Reynaud « Pour une souveraineté européenne sur le Cloud et les données », parue le 15 Novembre 2022 sur la plateforme La Grande Conversation de Terra Nova. Le constat est difficilement contestable : la dépendance Européenne envers les *cloud providers*, principalement américains - mais également chinois - s'accentue. Pour grand nombre d'entreprises ou administrations dont la donnée est au coeur des activités, cette dépendance technologique met en cause notre souveraineté dans toutes ses dimensions, pour toutes les raisons listées avec pertinence par Thomas Reynaud. Un cloud européen véritablement souverain s'impose, accompagné d'un *Buy European Act*.

Il m'apparait toutefois important de faire un pas de côté pour se pencher sur les tendances actuelles en traitement de données, afin de réfléchir au positionnement de ce futur cloud dans la chaîne de création de valeur. De fait, l'Intelligence Artificielle a plus évolué dans les 3 dernières années que dans toute la décennie précédente, avec l'apparition de méga-modèles (à plus de cent milliards de paramètres !) que l'on appelle les "Grands Modèles de Langage" (*Large Language Models*)**[1]**. Incroyablement puissants et versatiles, ces nouvelles IA effectuent de façon stupéfiante une série de tâches complexes sur le texte, comme la rédaction, le résumé, ou l'analyse sémantique. Ils sont maintenant étendus à des interfaces pour la génération *à volonté* d'images de synthèse d'une précision remarquable, heurtant des secteurs d'industries créatives s'étant crues à l'abri des technologies d'automatisation. Lancées par le modèle GPT-3 d'OpenAI (entreprise californienne massivement financée par Microsoft), des dizaines de startups se créent chaque semaine de par le monde pour exploiter ces modèles dans une grande diversité d'assistants IA, redoutablement efficaces pour une variété toujours plus grande de tâches du monde de l'entreprise, du travailleur indépendant à la multinationale**[2]**. Et nous n'en sommes qu'au début, avec une accélération proprement vertigineuse de la recherche dans ce domaine - recherche presque entièrement privée, notamment portée par des « GAFAM » en ébullition sur ces sujets -. Chaque semaine voit ainsi la naissance de nouveaux modèles, dont les plus avancés annoncent maintenant la prochaine génération, douée de capacités de planification et de raisonnement. Dès aujourd'hui, ils s'immiscent dans les suites bureautiques de Microsoft et Google – et sont donc susceptibles d'analyser toutes les données qui y transitent. Dans 1 ou 2 ans, il est probable qu'une technologie aussi centrale que les moteurs de recherche tels que nous les connaissons depuis vingt ans aient été remplacée par une nouvelle génération intégrant bien mieux le contexte de notre demande. Il ne s'agit pas de la science-fiction d'un HAL ou d'un Terminator, mais bien des prémisses d'une rupture technologique majeure, où l'IA va, pour le meilleur et pour le pire, bousculer toute la chaine de création de valeur dans des industries technologiques, mais aussi dans les services et les administrations. En d'autres termes, nous sommes au début d'une automatisation des « cols blancs », à tous les niveaux de responsabilité, du stagiaire au PDG. *Brace yourselves*.

[1] « Huge "foundation models" are turbo-charging AI progress », *The Economist*, 11 Juin 2022

[2] « AI's sudden big leap forward into usefulness », *Financial Times*, 6 Octobre 2022

Cette IA sur stéroïdes démultiplie l'intensité des enjeux mentionnés par Thomas Reynaud : d'une part l'absolue nécessité de mettre en place de meilleures règles de protection de la vie privée - alors que ces modèles ont une capacité accrue d'en croiser les sources -, ensuite un enjeu évidemment stratégique par rapport aux technologies de défense et de renseignement, et enfin des enjeux environnementaux, alors que les capacités de calcul nécessaires pour faire tourner ces méga-modèles sont sans commune mesure avec la génération précédente et qu'il s'agira de *scaler* l'utilisation quotidienne de tels modèles à l'échelle de milliards d'utilisateurs.

C'est là où j'apporterais une nuance à l'affirmation de Thomas Reynaud: certes les Data Centers ne suffisent pas, mais dans ce nouveau contexte tout matériel de calcul ne se vaut pas, à l'inverse d'une *commodity*. Il est au contraire essentiel que la couche physique, les serveurs de calcul eux-mêmes, soient en capacité de faire tourner l'IA la plus puissante. A l'heure actuelle ce matériel a un nom, avec un quasi-monopole *de facto* des cartes GPU (Graphical Processing Unit) de NVIDIA, dont les modèles les plus puissants**[3]** sont conçus explicitement pour cet usage, et de fait extrêmement chers et gourmands en énergie. Accorderais-je une importance exagérée à ces processeurs ? Ils sont tellement d'une autre nature que les Etats-Unis ont récemment étendu leur réglementation d'*export control* spécifiquement sur ces modèles de cartes**[4]**, maintenant interdits d'export vers notamment la Chine, au même titre donc que des matériels explicitement à usage militaire ou de renseignement.
Or, dans l'offre actuelle des cloud providers européens, aucune de ces cartes haut de gamme pourtant disponibles depuis 2 ans - une éternité ! - chez les « Big 3 » (AWS - Amazon -, GCP – Google/Alphabet -, Azure - Microsoft -). Impossible pour une société comme la mienne d'y faire tourner nos modèles - installés donc, pour nos clients, sur un *hyperscaler* américain extrêmement efficace. L'Etat français a financé - pour un équivalent de plusieurs millions d'euros - l'entrainement du modèle BLOOM open-source sur le super calculateur de recherche "Jean Zay" - effort de recherche communautaire remarquable auquel plus de 500 chercheurs ont contribué -, mais ce modèle à 176 milliards de paramètres, maintenant entraîné, est trop gros pour tourner sur les *cloud providers* français et véritablement bénéficier à l'écosystème**[5]**.
L'offre actuelle réduite aux serveurs de milieu de gamme se justifie parfaitement dans la logique économique rappelée par Thomas Reynaud des 20/80 : 20% de l'offre pour couvrir 80% des besoins - dans le contexte extrêmement concurrentiel qui est souligné dans sa note. Mais, dans la perspective d'un Cloud souverain européen il serait incompréhensible d'avoir pour stratégie de dupliquer ce modèle tout en évitant les errements de GAIA-X, et de fait exclure les technologies de l'IA les plus avancées, à très forte valeur ajoutée.

La bonne nouvelle est que cette nécessaire montée en gamme est techniquement parfaitement faisable, car il s'agit avant tout de matériel (« hardware »). Certes, le surcoût d'infrastructure est significatif, mais la demande croissante saura s'ajuster aux prix. C'est également à cette aune que sera jugée la crédibilité d'un positionnement véritablement souverain d'un *cloud*, capable de répondre à des usages parmi les plus structurants pour le futur de nos sociétés.

---

[3] Cartes GPU NVIDIA A100 (génération 2020) et H100 (génération 2022)

[4] https://www.reuters.com/technology/nvidia-says-us-has-imposed-new-license-requirement-future-exports-china-2022-08-31/

[5] « Intelligence artificielle : les multiples visages du deep learning », *Le Monde*, 24 octobre 2022