Into the cloud… inspiré d’une histoire vraie

Spécialiser la traduction automatique en ligne avec l’intelligence artificielle

OVHcloud

SYSTRAN
beyond language

+ de 30 000 heures
cumulées de calcul sur des GPU NVIDIA Tesla V100

50 000 modèles
intermédiaires stockés dans l'Object Storage pour un volume de 30 Tbit

Une capacité de traduction de **5 milliards de mots** par jour

Le contexte

SYSTRAN est un éditeur de solutions professionnelles de traduction automatique, qui a fêté ses 50 ans en 2018.

Avec plus de 140 paires de langues disponibles, les services de SYSTRAN sont personnalisés selon chaque contexte client. Ils sont utilisés par de nombreuses entreprises internationales, organisations publiques et agences de traduction.

Depuis sa création, SYSTRAN a toujours été pionnier dans le traitement automatique des langues. Fin 2016, l'entreprise a été de nouveau précurseur en lançant le premier moteur de traduction neuronale professionnel. Celui-ci tire parti des dernières avancées offertes par les technologies de deep learning, afin d'améliorer la qualité de la traduction instantanée.

Client OVHcloud de la première heure, l'éditeur s'est associé au fournisseur en 2018 pour élaborer une offre baptisée SYSTRAN Marketplace. Cette plateforme communautaire permet de proposer les meilleurs modèles de traduction du marché, entraînés par des experts multilingues de différents domaines. Disponibles dans le cloud ou on-premises, via des outils de traduction professionnels, les modèles sont intégrés dans le système d'information du client.

Pour relever ce défi, SYSTRAN a choisi une approche communautaire basée sur quatre piliers : la technologie, les données, l'expertise humaine et l'infrastructure pour offrir une solution ouverte, responsable, dimensionnée pour le web et hautement disponible.

Le défi

Depuis 2016, le monde de la traduction automatique a considérablement évolué. La traduction neuronale – une approche issue de la recherche en intelligence artificielle et, en particulier, du deep learning – s'est imposée comme le standard, succédant à la traduction dite statistique. Cette dernière était essentiellement basée sur le big data, ainsi que sur la représentation par des experts des règles gouvernant les langues.

Des changements profonds ont accompagné cette transition. Sur le plan technologique, les algorithmes nécessaires ne cessent d'évoluer et sortent directement des grands laboratoires de recherche privés et publics. Grâce à l'approche neuronale, un courant open source général s'est développé et imposé, permettant ainsi une progression scientifique reproductible et un développement industriel quasiment instantané.

Si la quantité de données nécessaire est moins importante qu'auparavant, la qualité de celles-ci est primordiale tant les modèles neuronaux vont tenter d'interpréter tout « bruit » comme des règles de langue. Le big data fait oublier que les informations utilisées pour entraîner des modèles de traduction sont produites par des traducteurs humains. Or, si ces données semblent disponibles en ligne, elles n'en sont pas moins soumises au droit d'auteur. Et la qualité d'un modèle résulte directement de l'investissement dans ces mêmes données, ce qui impose une parfaite traçabilité. Sans cette rigueur, il serait dangereux de faire confiance à des modèles de traduction qui pourraient avoir été biaisés par leurs informations source.

L'expertise humaine, mise de côté à l'époque statistique, revient également en force. Si les algorithmes sont extrêmement puissants, ils ont besoin d'être supervisés par des spécialistes de la langue et de différents domaines.

Enfin, l'approche neuronale a profondément changé les besoins pour les infrastructures de calcul. Pendant la phase d'entraînement des modèles, comme pour tout algorithme de deep learning, des cartes graphiques (GPU) spécifiques sont nécessaires. En revanche, lors de l'inférence, c'est-à-dire l'utilisation des modèles en production, les algorithmes nécessitent des serveurs optimisés pour le calcul et relativement peu de mémoire en comparaison avec les générations précédentes. Aussi l'évolution de la réglementation pour mieux protéger les droits des utilisateurs implique une attention particulière sur les infrastructures hébergeant des services pouvant traduire des données confidentielles.

Au-delà de l'apparente simplicité liée à chacun de ces changements – souvent illustrée par des démonstrations de performance dans des cas d'utilisation extrêmement restreints – des modifications fondamentales sont nécessaires, afin de fournir une chaîne de production à grande échelle à la fois responsable, transparente et capable de fournir la meilleure qualité pour chaque industrie. Le principe fondamental de cette approche consiste à reconnaître l'expertise des différents acteurs présents et de les associer pour atteindre l'excellence.

De son côté, SYSTRAN a d'abord investi dans l'open source en cofondant dès 2016 OpenNMT, un framework d'algorithmes de traduction neuronale. Cette technologie, aujourd'hui la plus populaire et active dans son secteur, est utilisée par des milliers de chercheurs et industriels, qui l'enrichissent quotidiennement avec leurs contributions. Grâce à cette brique logicielle de pointe, les équipes R&D de SYSTRAN ont conçu des solutions complètes de traduction pensées pour les utilisateurs finaux. Enfin, l'éditeur a développé une marketplace composée de plusieurs services. Celle-ci permet à une communauté d'experts de produire et partager des modèles de haute qualité, tout en étant directement rémunérés pour leurs contributions.

Afin de bâtir cette plateforme, le choix d'une infrastructure flexible, robuste et adaptable s'est vite imposé. Elle devait offrir la puissance de calcul nécessaire à l'entraînement des moteurs neuronaux. Cet environnement devait également être évolutif pour déployer ses modèles en production, répondre aux fluctuations des demandes, ainsi que respecter l'esprit responsable de cette approche communautaire… le tout à un prix compétitif.

La solution

Une plateforme ouverte, sécurisée et responsable, parfaitement adaptée aux besoins du deep learning

> *« Le choix d'OVHcloud comme partenaire technologique pour l'hébergement et l'exploitation de notre marketplace s'est imposé rapidement. L'ADN même d'OVHcloud correspondait à l'esprit de la marketplace. Nos exigences en matière de flexibilité et de puissance nous ont directement conduits vers l'offre Public Cloud. »*
>
> **Jean Senellart, Président Directeur Général de SYSTRAN**

Une solution technique combinant puissance, flexibilité et prédictibilité

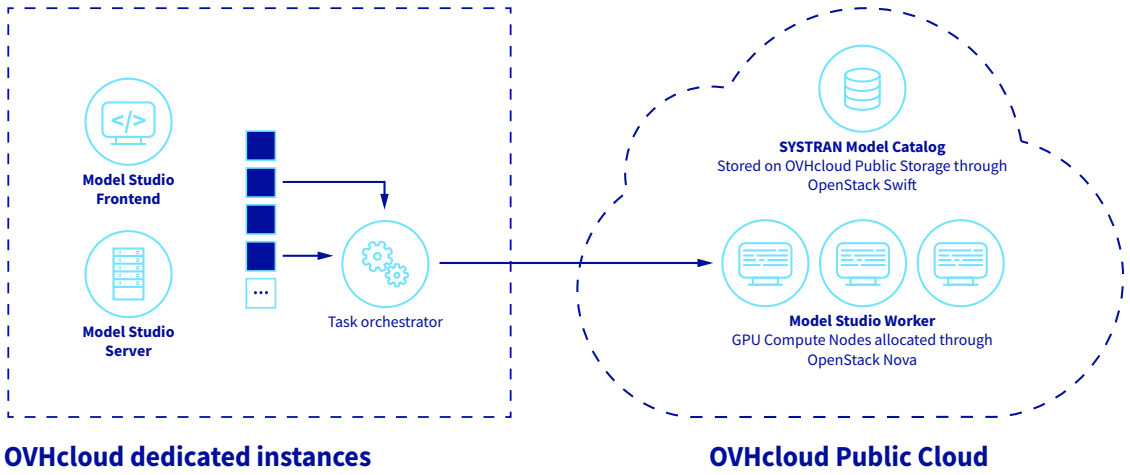
Afin de mener à bien ce projet, SYSTRAN a opté pour la solution Public Cloud. En effet, cette dernière permet une pleine maîtrise des coûts, avec un accès à une large gamme de serveurs et de services. Elle offre aussi la flexibilité nécessaire pour entraîner des modèles neuronaux à la demande et gérer des volumes de traduction variant au fil du temps.

SYSTRAN Model Studio – une solution unique développée par SYSTRAN pour permettre à des experts de la langue et d'un domaine d'entrainer eux-mêmes leurs modèles de traduction – a besoin d'accéder à la demande aux processeurs graphiques (GPU) les plus puissants du marché. La contrainte n'était pas ici la disponibilité instantanée des instances de calcul, car l'entraînement de modèles neuronaux repose sur des cycles allant de quelques heures à une semaine.

Model Studio est un orchestrateur de tâches, capable de gérer une séquence d'itérations correspondant à un entraînement donné. Il utilise l'API Nova d'OpenStack afin de lancer des instances de calcul dynamiquement.

Dans ce schéma, la fiabilité des instances est essentielle. En effet, une itération qui échouerait aurait pour conséquence l'échec de l'entraînement associé et la perte de journées de calcul.

Model Studio nécessite aussi une énorme capacité de stockage, puisque chaque itération d'un entraînement est un réseau de neurones archivé et testé. À savoir qu'un modèle représente des milliards de paramètres, soit plusieurs gigaoctets, qui sont stockés sur l'Object Storage via le service Swift d'OpenStack organisé en conteneurs.



Cette infrastructure a été mise au point en un an. Durant cette période, les équipes de SYSTRAN ont pu entraîner des centaines de modèles en utilisant un pool basé sur des serveurs NVIDIA DGX-1, ainsi que des pools complémentaires de Public Cloud s'appuyant sur des instances GPU NVIDIA Tesla V100. La plateforme est maintenant à la disposition des « entraineurs » de la marketplace, pour leur permettre de créer leurs propres modèles en toute autonomie.

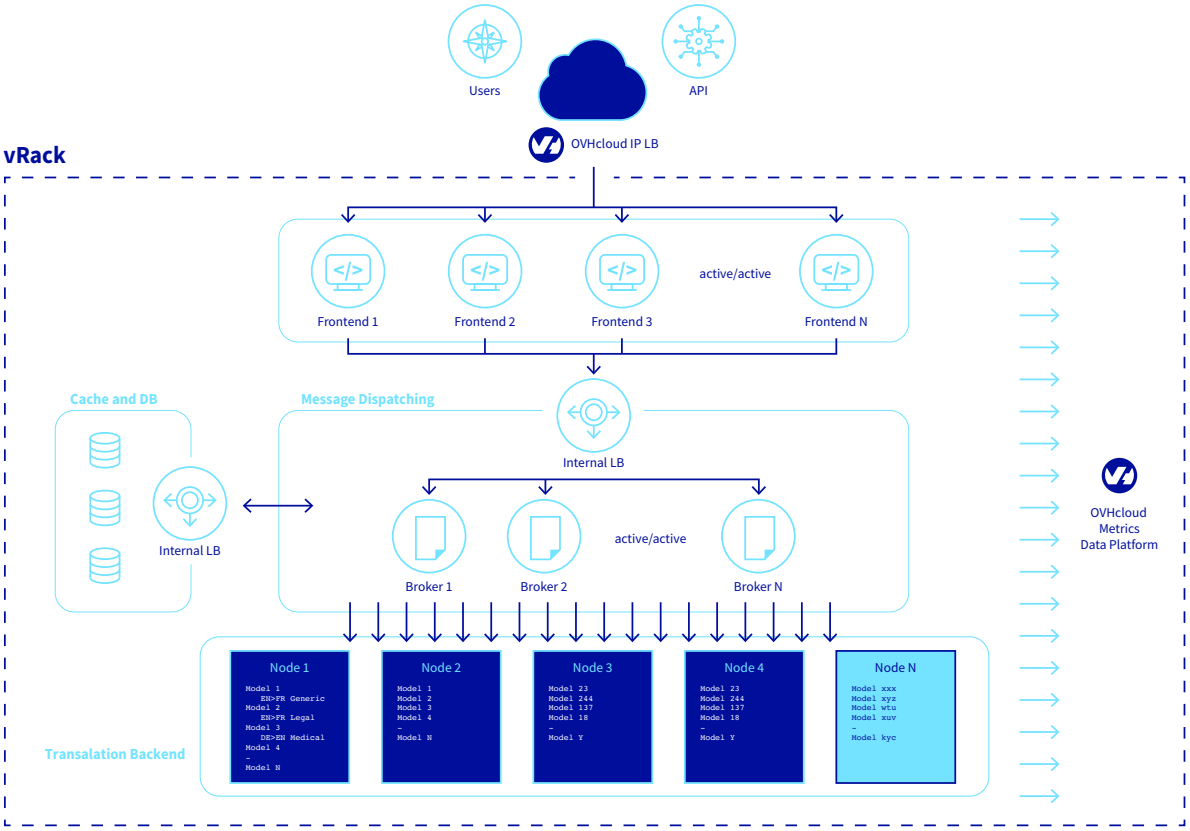
Pour l'inférence, la problématique est inverse. Le service doit être disponible 24 heures/24 et s'adapter à la volumétrie des requêtes à un instant t, tout en utilisant des instances optimisées pour le calcul. Aussi, chaque requête doit être traitée en quelques millisecondes et requiert un mix d'instances statiques et dynamiques.

Le point d'entrée de l'infrastructure de la plateforme SYSTRAN Translate est un load balancer. Son rôle est crucial, puisqu'il répartit la charge entre les différents services hébergés dans les datacenters et protège l'application contre les attaques DDoS. Cet équipement assure aussi la mise à l'échelle de l'infrastructure en cas de pic de trafic. Enfin, il permet de garantir une haute disponibilité de service et d'optimiser les temps de réponse.

En juillet 2019, l'infrastructure était composée de 74 instances Public Cloud GPU. Celle-ci est sécurisée par le vRack, une interconnexion privée *made in* OVHcloud.

Pour aller plus loin, les équipes ont ajouté une composante dynamique au service. Basée sur Kubernetes, elle permet d'allier disponibilité instantanée et dimensionnement flexible de l'infrastructure.

Cette dernière est monitorée par la plateforme managée Metrics Data Platform. Cela assure un suivi en temps réel de chacun des composants, mais aussi des temps de réponse et des volumes de traduction pour toutes les paires de langues et modèles.



Une plateforme basée sur des standards ouverts

Le développement de l'ensemble de l'infrastructure de la marketplace a été grandement facilité grâce aux services d'OVHcloud. Tous équipés d'API open source, ils garantissent une prise en main immédiate par les équipes de développement.

> *« Le choix et l'investissement dans des solutions open source garantissent simultanément aux utilisateurs finaux le meilleur de la technologie disponible et aux développeurs ainsi qu'aux contributeurs de la marketplace de ne pas se retrouver enfermés par des technologies propriétaires. »*
>
> **Yannick Douzant, Directeur Produits et Technologies de SYSTRAN**

Pour SYSTRAN, qui développe et maintient l'intégralité du code de traduction neuronale dans le projet OpenNMT, comme pour OVHcloud, qui a fait le choix de nombreux standards ouverts pour son service Public Cloud, cette approche open source, au-delà de sa facilité de prise en main, est une composante importante de la philosophie autour du développement logiciel commune aux deux entreprises.

Une approche responsable

> *« L'engagement d'OVHcloud pour une écoresponsabilité dans la conception des serveurs, dans l'opération avec un système exclusif de watercooling, dans une politique développant une énergie verte et dans le recyclage des composants en fin de vie des équipements a été un critère de choix prépondérant pour l'infrastructure de notre marketplace. »*
>
> **Jean Senellart, Président Directeur Général de SYSTRAN**

Quant aux données, elles sont protégées et assurées de ne pas quitter le sol européen pour assurer le respect du règlement général sur la protection des données (RGPD).

Le résultat

Grâce à la technologie utilisée et à l'accompagnement des experts d'OVHcloud, il n'aura fallu que deux semaines aux équipes techniques de SYSTRAN pour déployer et mettre en ligne le service SYSTRAN Translate.

Cinq mois seulement après son lancement, le service a déjà permis à plus d'un million d'utilisateurs provenant de 190 pays de traduire des milliards de mots. Il bénéficie d'une forte popularité en Europe et, en particulier, en France, au Royaume-Uni, en Belgique et en Allemagne.

Le service de traduction automatique propose plus de 40 langues et met à disposition 400 modèles de traduction. D'ici un an, l'objectif est d'atteindre 5 000 modèles grâce à l'expansion de la communauté d'experts.

Ce n'est qu'un début, puisque SYSTRAN Translate ne représente que la première brique d'une nouvelle offre destinée aux professionnels : SYSTRAN Marketplace. Celle-ci a pour ambition de leur proposer le plus large catalogue de modèles spécialisés, assorti de la plus large gamme de solutions de traduction déployées on-premises ou dans le Cloud, en mode privé ou public. Pour répondre à tous les types de besoins et à tous les volumes, avec le même niveau de qualité.