"Into the cloud"... basato su una storia vera

Perfezionare la traduzione automatica online con l'intelligenza artificiale

OVHcloud

SYSTRAN
beyond language

+ di 30.000 ore
di calcolo accumulate
su GPU Nvidia Tesla V100

50.000 modelli
intermedi archiviati
nell'Object Storage
per un volume di 30 Tb

Capacità di traduzione
di **5 miliardi di parole**
al giorno

Executive Summary

SYSTRAN è un'azienda di soluzioni professionali di traduzione automatica che nel 2018 ha festeggiato i primi 50 anni di vita.

Con oltre 140 combinazioni linguistiche disponibili, i suoi servizi sono personalizzabili in base al contesto di ciascun cliente e vengono utilizzati da numerose aziende internazionali, organismi pubblici e agenzie di traduzione.

Da quando è stata fondata, SYSTRAN è un pioniere nel trattamento automatico delle lingue. Alla fine del 2016 si è di riproposta come avanguardia del settore lanciando il primo motore professionale di traduzione neurale, sviluppato grazie agli ultimi progressi delle tecnologie di Deep Learning volti a migliorare la qualità delle traduzioni istantanee.

Cliente OVHcloud sin dagli esordi, nel 2018 l'azienda si è associata al provider per elaborare il servizio SYSTRAN Marketplace, una piattaforma comunitaria che fornisce i migliori modelli di traduzione del mercato istruiti da esperti linguistici di diversi settori. Questi modelli, disponibili sia nel Cloud che in locale tramite strumenti di traduzione professionali, sono integrati nel sistema IT del cliente.

Per far fronte alla sfida, SYSTRAN ha optato per un approccio comunitario basato su quattro pilastri: tecnologia, dati, competenze umane e infrastruttura. L'obiettivo era offrire una soluzione aperta, responsabile, disegnata per il Web e dalla massima disponibilità.

La Sfida

Dal 2016 il mondo della traduzione automatica ha conosciuto un'evoluzione significativa. La traduzione neurale – un approccio scaturito dalla ricerca nell'ambito dell'Intelligenza Artificiale e, in particolare, del Deep Learning – si è imposta come standard, subentrando alla traduzione detta statistica basata essenzialmente sul Big Data e sulla rappresentazione delle regole linguistiche da parte degli esperti.

Questa transizione è stata accompagnata da profondi cambiamenti. Dal punto di vista tecnologico, gli algoritmi sono in continua evoluzione e provengono direttamente dai grandi laboratori di ricerca privati e pubblici. Grazie all'approccio neurale, si è sviluppata e imposta una corrente open source generale che consente un progresso scientifico riproducibile e uno sviluppo industriale quasi istantaneo.

Se la quantità di dati necessaria è meno importante rispetto al passato, la qualità è invece essenziale per i modelli neurali che tentano di interpretare qualsiasi input come una regola linguistica. Il Big Data porta spesso a dimenticare che le informazioni utilizzate per l'apprendimento dei modelli di traduzione sono il prodotto di traduttori umani e che, anche se disponibili online, potrebbero essere vincolate al diritto d'autore. La qualità dei modelli, inoltre, è una conseguenza diretta dell'investimento in questi stessi dati, che richiede una perfetta tracciabilità. Senza questo rigore fidarsi di modelli di traduzione sarebbe pericoloso, perché potrebbero includere contenuti distorti rispetto ai testi originali.

Le competenze umane, passate in secondo piano nell'epoca statistica, ritornano così in auge. Nonostante gli algoritmi siano estremamente potenti, infatti, richiedono una supervisione da parte di specialisti linguistici di diversi settori.

L’approccio neurale ha profondamente cambiato anche le esigenze in termini di infrastrutture di calcolo: come per qualsiasi algoritmo di Deep Learning, nella fase di addestramento dei modelli sono necessarie schede grafiche (GPU) specifiche. Durante l’inferenza invece, cioè l’utilizzo dei modelli in ambienti di produzione, gli algoritmi richiedono server ottimizzati per il calcolo e una capacità di memoria relativamente ridotta rispetto alle generazioni precedenti. Anche l’evoluzione della legislazione relativa alla protezione dei diritti degli utenti impone un’attenzione particolare alle infrastrutture che ospitano servizi che possono potenzialmente tradurre dati confidenziali.

Al di là dell’apparente semplicità legata a queste evoluzioni – spesso illustrata da dimostrazioni di performance in casi di utilizzo estremamente limitati – sono necessarie modifiche fondamentali, per fornire una catena di produzione su larga scala responsabile, trasparente e in grado di fornire la qualità migliore in tutti i settori. Questo approccio si basa su un principio di fondo: riconoscere l’importanza dei diversi soggetti coinvolti e associarli per raggiungere l’eccellenza.

Dal canto suo, SYSTRAN ha per prima cosa investito nell’open source cofondando nel 2016 OpenNMT, un framework di algoritmi di traduzione neurale. Questa tecnologia, attualmente la più diffusa e attiva del settore, è utilizzata da migliaia di ricercatori e industriali che la arricchiscono ogni giorno con i propri contributi. Grazie a questo software di punta, i team R&D di SYSTRAN hanno progettato soluzioni di traduzione complete, pensate per gli utenti finali. L’azienda ha inoltre sviluppato un marketplace composto da diversi servizi che permette a una Community di esperti di produrre e condividere modelli di alta qualità ed essere remunerati per il contributo offerto.

Per costruire questa piattaforma era indispensabile un’infrastruttura flessibile, solida e adattabile che offrisse una potenza di calcolo adeguata per l’addestramento dei motori neurali. Questo ambiente doveva anche essere scalabile per poter implementare i modelli in ambienti di produzione, rispondere alle fluttuazioni del numero di richieste e rispettare lo spirito responsabile dell’approccio comunitario... il tutto a tariffe competitive.

La Soluzione

Una piattaforma aperta, sicura e responsabile ottimizzata per le esigenze in termini di Deep Learning.

> *"La scelta di OVHcloud come partner tecnologico per l'hosting e l'implementazione del nostri marketplace si è imposta rapidamente. I principi radicati nel DNA di OVHcloud rispecchiano lo spirito del marketplace. La nostra necessità di flessibilità e potenza ci ha condotti direttamente verso la soluzione Public Cloud."*
>
> **Jean Senellart, CEO di SYSTRAN**

Una soluzione tecnica che combina potenza, flessibilità e prevedibilità.

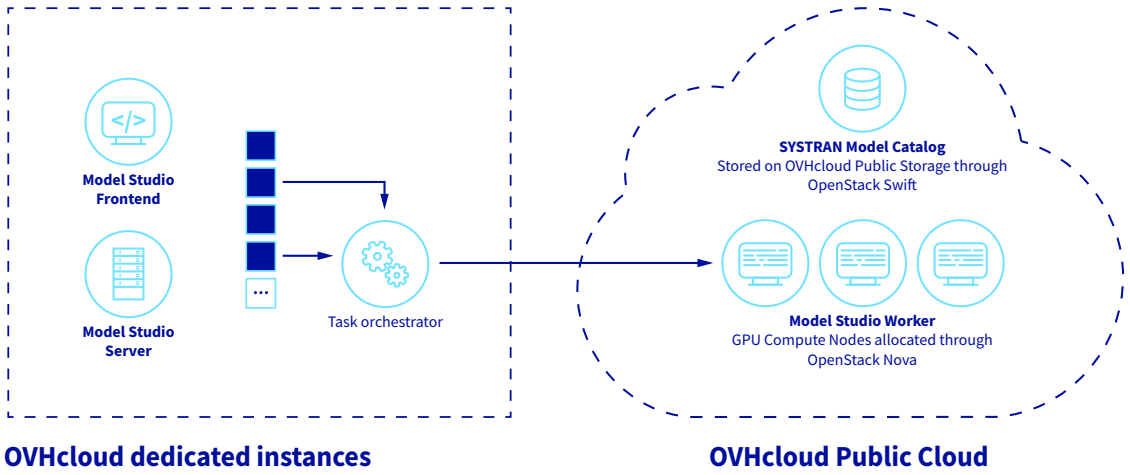
Per portare a termine il progetto SYSTRAN ha optato per la soluzione Public Cloud, che permette un controllo totale dei costi e l'accesso a una vasta gamma di server e servizi. Offre anche la flessibilità necessaria per istruire i modelli neurali on demand e gestire volumi di traduzione variabili nel tempo.

SYSTRAN Model Studio – una soluzione unica sviluppata da SYSTRAN per permettere agli esperti linguistici settoriali di istruire in autonomia i propri modelli di traduzione – richiede l'accesso on demand ai processori grafici (GPU) più potenti del mercato. In questo caso il vincolo non era la disponibilità istantanea delle istanze di calcolo, perché l'istruzione dei modelli neurali si basa su cicli che variano da un'ora a una settimana.

Model Studio è un orchestratore di task in grado di gestire una sequenza di iterazioni che corrispondono a un addestramento stabilito. Per avviare istanze di calcolo in modo dinamico, utilizza l'API Nova di OpenStack.

In questo contesto, l'affidabilità delle istanze è fondamentale: un'iterazione non andata a buon fine avrebbe infatti come conseguenza il fallimento dell'addestramento associato e la perdita di giornate di calcolo.

Model Studio richiede anche un'enorme capacità di storage, in quanto ogni iterazione di un addestramento è una rete neurale archiviata e testata. Un modello è costituito infatti da miliardi di parametri, cioè diversi GB memorizzati nell'Object Storage tramite il servizio Swift di OpenStack organizzato in container.



Questa architettura è stata realizzata in un anno, nel corso del quale i team di SYSTRAN hanno potuto istruire centinaia di modelli utilizzando server NVIDIA DGX-1 e altri pool complementari Public Cloud basati su istanze GPU NVIDIA Tesla V100. La piattaforma è ora a disposizione degli "allenatori" del marketplace, che hanno la possibilità di creare i propri modelli in completa autonomia.

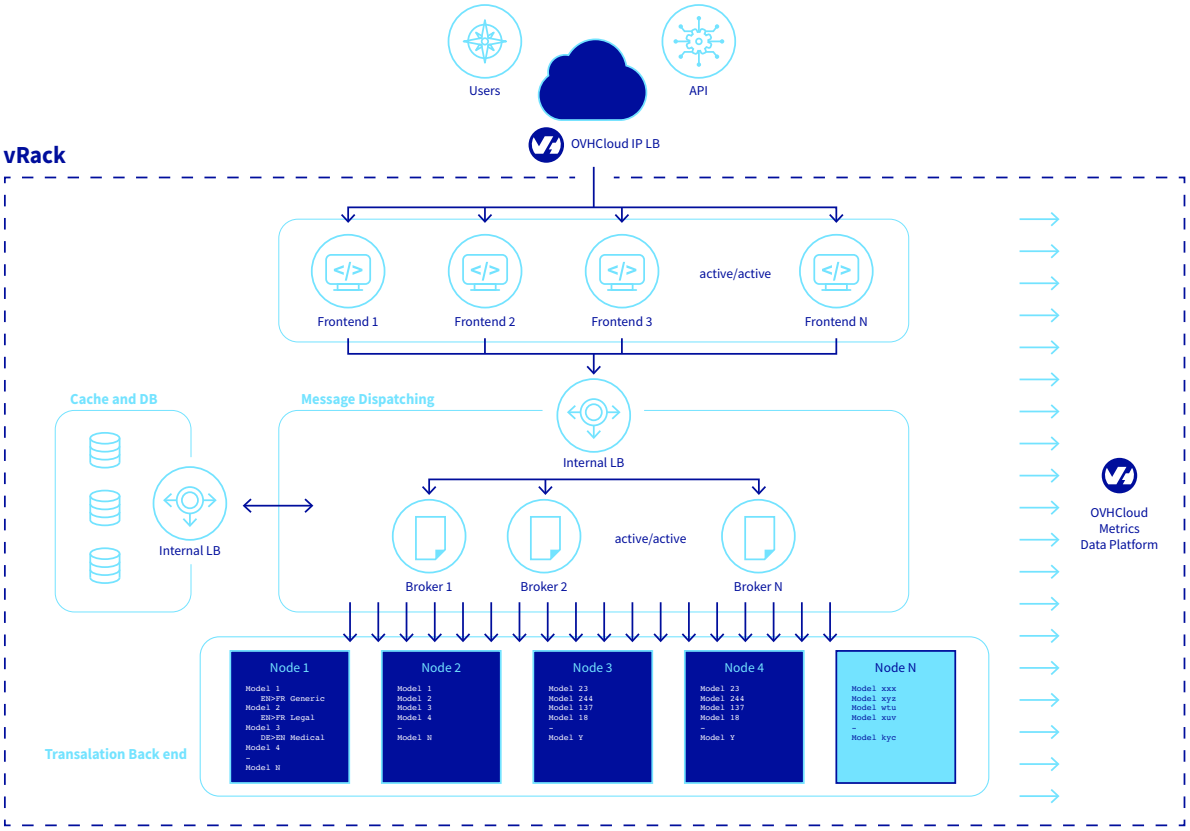
Per l'inferenza, la problematica è inversa: il servizio deve essere disponibile 24 ore su 24 e adattarsi al volume di richieste in un istante T, utilizzando istanze ottimizzate per il calcolo. Ogni richiesta deve inoltre essere elaborata in pochi millisecondi e richiede un mix di istanze statiche e dinamiche.

Il punto di entrata dell'infrastruttura della piattaforma SYSTRAN Translate è un Load Balancer, il cui ruolo è cruciale in quanto distribuisce il carico tra i diversi servizi ospitati nei datacenter e protegge l'applicazione dagli attacchi DDoS. Questo dispositivo assicura anche lo scaling dell'infrastruttura in caso di picchi di traffico, permette di garantire la massima disponibilità del servizio e di ottimizzare i tempi di risposta.

A luglio 2019 l'infrastruttura era composta da 74 istanze Public Cloud GPU e protetta tramite la vRack, un'interconnessione privata made in OVHcloud.

I team hanno deciso di spingersi oltre aggiungendo una componente dinamica al servizio. Basato su Kubernetes, questo elemento permette di combinare disponibilità istantanea e dimensionamento flessibile dell'infrastruttura.

Il monitoraggio avviene tramite la piattaforma gestita Metrics Data Platform, che assicura la supervisione in tempo reale di singoli componenti, tempi di risposta e volumi di traduzione per tutte le combinazioni linguistiche e modelli.



OVHcloud

Una piattaforma basata su standard aperti.

Lo sviluppo dell'intera infrastruttura del marketplace è stato enormemente facilitato dai servizi di OVHcloud: tutti dotati di API open source, garantiscono un'elaborazione immediata delle richieste da parte dei team di sviluppo.

> *"La scelta e l'investimento in soluzioni open source garantiscono agli utenti finali il meglio della tecnologia disponibile ed evitano a sviluppatori e contributor della Community di ritrovarsi vincolati a tecnologie proprietarie."*
>
> **Yannick Douzant, Responsabile Prodotti e Tecnologie di SYSTRAN**

Sia per SYSTRAN, che sviluppa e mantiene l'integralità del codice di traduzione neurale nel progetto OpenNMT, che per OVHcloud, che ha scelto numerosi standard aperti per il proprio servizio Public Cloud, l'approccio open source va oltre la semplicità di utilizzo e rappresenta un fattore importante della filosofia che ruota attorno allo sviluppo software comune alle due aziende.

Un approccio responsabile.

> *"L'impegno di OVHcloud verso l'ecoresponsabilità nella concezione dei server, nell'utilizzo del sistema esclusivo di watercooling, nella politica dello sviluppo di un'energia verde e nel riutilizzo dei componenti alla fine del loro ciclo di vita è stato un criterio preponderante nella scelta dell'infrastruttura del nostro marketplace."*
>
> **Jean Senellart, CEO di SYSTRAN**

Quanto ai dati, sono protetti e legati al suolo europeo per garantire il rispetto del Regolamento Generale sulla protezione dei dati (GDPR).

I Risultati

Grazie alla tecnologia utilizzata e all'accompagnamento degli esperti OVHcloud, i team tecnici di SYSTRAN hanno impiegato solo due settimane per implementare e pubblicare SYSTRAN Translate.

Solo cinque mesi dopo il lancio, questo servizio di traduzione automatica aveva già consentito a oltre un milione di utenti provenienti da 190 Paesi di tradurre miliardi di parole. Molto diffuso in Europa, in particolare in Francia, Regno Unito, Belgio e Germania, propone oltre 40 lingue e mette a disposizione 400 modelli di traduzione. L'obiettivo è raggiungere entro un anno i 5.000 modelli, grazie all'espansione della Community di esperti.

Non è che l'inizio, perché SYSTRAN Translate rappresenta il primo livello di una nuova offerta destinata ai professionisti: SYSTRAN Marketplace. Questa nuova soluzione ha l'ambizione di proporre il più ampio catalogo di modelli specializzati, accompagnato dalla più vasta gamma di servizi di traduzione implementati in locale o nel Cloud, in modo privato o pubblico. Per rispondere a qualsiasi tipo di esigenza offrendo lo stesso livello di qualità.

OVHcloud è un provider Cloud globale specializzato nell'offerta di soluzioni competitive e con prestazioni di alto livello per gestire, proteggere e scalare i dati nel modo migliore. OVHcloud rappresenta l'alternativa più intelligente per soluzioni di hosting Web, email, server bare metal, Hosted Private Cloud, Public Cloud e hybrid Cloud. Il gruppo gestisce 30 datacenter nelle 12 localizzazioni distribuite in 4 continenti, assemblando i propri server, costruendo i propri datacenter e implementando la propria rete globale in fibra ottica per ottenere la massima efficienza. Grazie al suo spirito di sfida allo status quo OVHcloud porta libertà, sicurezza e innovazione per risolvere le prove attuali e future legate ai dati. Con un patrimonio ventennale e solide fondamenta in Europa, l'azienda lavora allo sviluppo di tecnologie responsabili e lotta per essere la forza motrice della prossima evoluzione del Cloud.