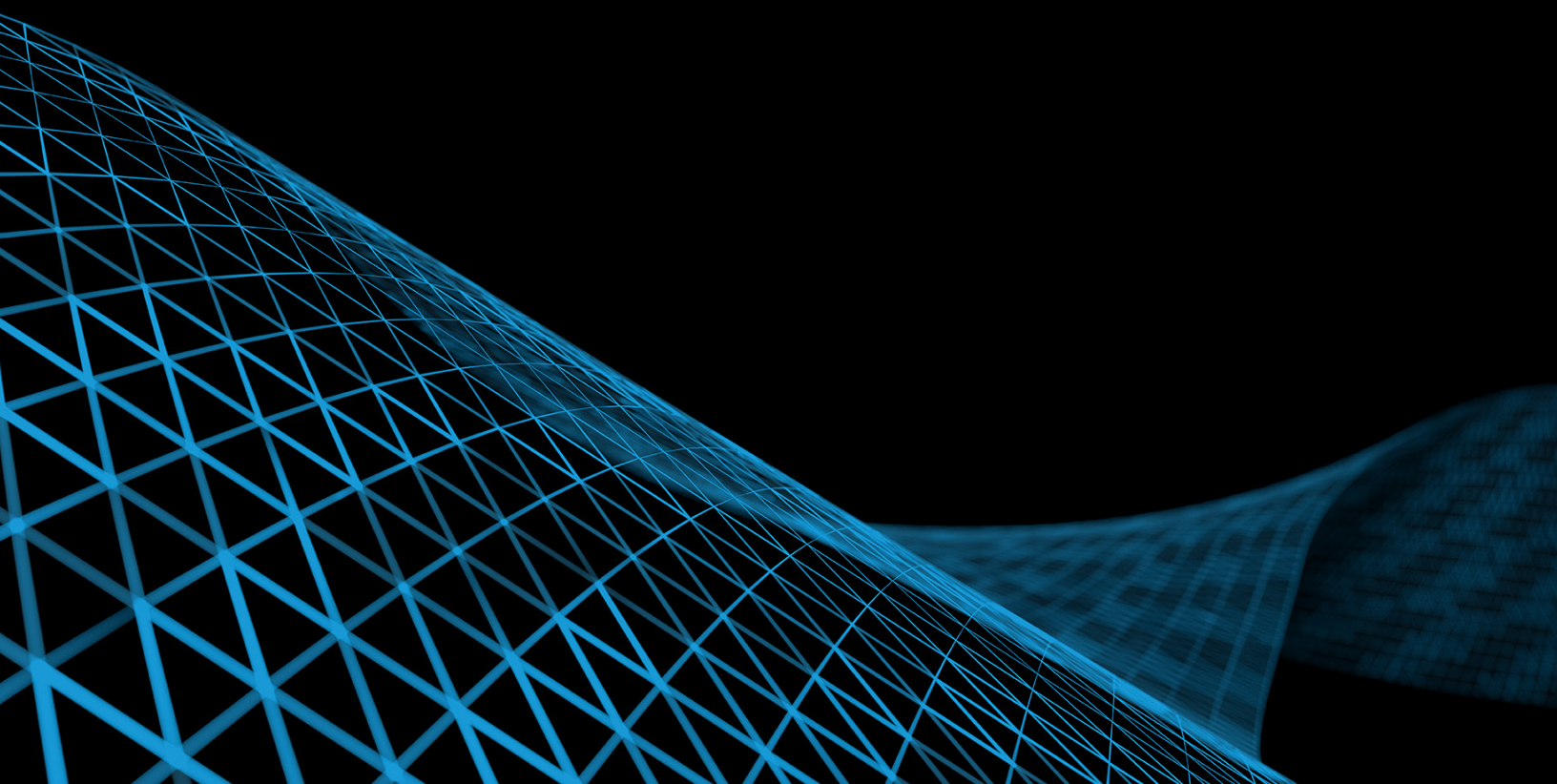
TechTarget | Custom Media
E-Book

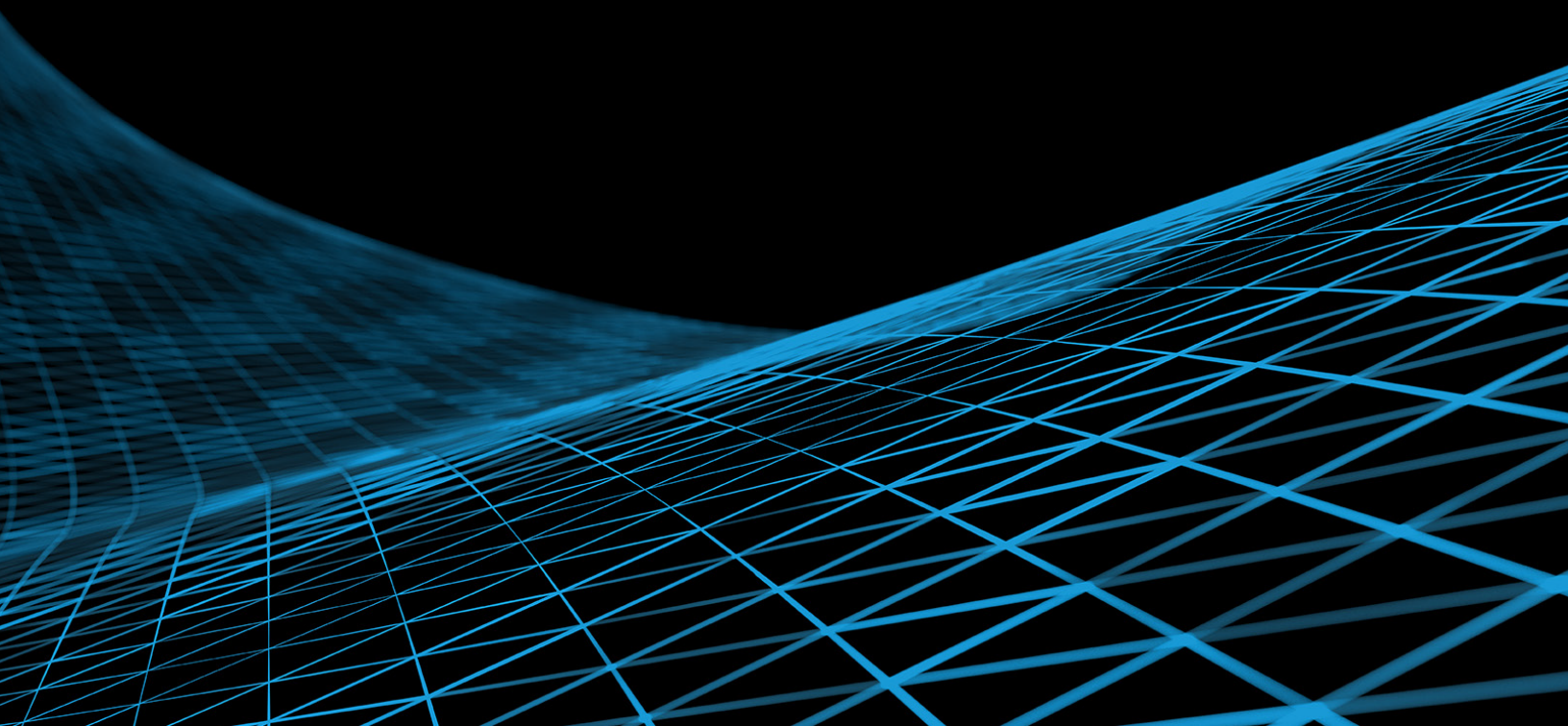
Come l'AI sta cambiando tutto, dal data center all'edge fino al cloud

Non si rimarcherà mai abbastanza l'impatto dell'AI sui processi di trasformazione in atto. L'intelligenza artificiale sta cambiando proprio il modo con cui il cloud e l'Edge Computing apportano valore alle organizzazioni in tutti i settori, le aree geografiche e i casi di utilizzo. Questo eBook offre una prospettiva pratica sul ruolo sempre più importante svolto dall'AI nel cloud distribuito.

Akamai



Sommario



```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Introduzione

Perché questo eBook è importante

I responsabili decisionali IT, i dirigenti di primo livello, i proprietari di segmenti aziendali o i membri del consiglio di amministrazione devono, probabilmente, impiegare sempre più tempo a leggere, parlare e pensare all'impatto dell'AI sulle loro organizzazioni e, molto probabilmente, sia da un punto di vista personale che professionale.

L'AI ha intrapreso il suo percorso di sviluppo sul mercato molto più rapidamente rispetto alla maggior parte delle tecnologie più innovative che l'hanno preceduta. L'AI è una delle più importanti tecnologie degli ultimi anni, che ha dato effettivamente ragione al clamore suscitato all'inizio nel settore e che ha fatto registrare risultati tangibili e concreti (più avanti, riporteremo alcuni dati significativi per illustrare questo importante fenomeno).

Questo eBook, tuttavia, non tratta solo della crescita dell'AI e del suo potenziale all'interno di un contesto realistico, ma esamina l'AI da un esclusivo punto di vista: l'AI vicina agli utenti e nel cloud distribuito. Abbiamo sentito parlare moltissimo dell'importanza di disporre di dati vicini agli utenti e nel cloud, ma ora, siamo entrati nell'epoca dell'AI inferencing decentralizzato e dell'AI nel cloud distribuito,

il che ci riporta al nostro titolo: "Perché questo eBook è importante". Di seguito, riportiamo quattro motivi per cui consigliamo la lettura di questo eBook a dirigenti/responsabili decisionali e addetti alla sicurezza:

1. **L'Edge Computing e il cloud computing sono le nuove frontiere dell'innovazione dell'AI e del valore che generano.** Creando e implementando le funzionalità dell'AI all'esterno del data center, le organizzazioni possono trarre vantaggio dalle nuove informazioni, dai modelli aziendali e dalle soluzioni che seguono lo stesso percorso intrapreso dal più ampio mercato IT nell'ultimo decennio.
2. **L'AI si basa su un colosso di normative e governance.** Se pensate che l'utilizzo e le pratiche "responsabili" dell'AI possano creare problemi se l'AI è stata implementata perlopiù nelle sandbox on-premise, considerate i problemi di conformità, protezione dei dati e sicurezza potenzialmente causati dall'AI vicina agli utenti e nel cloud distribuito. La vostra organizzazione deve affrontare questo aspetto con accuratezza o rischia di incorrere in tantissimi problemi.
3. **L'impatto dell'AI sui modelli e sulle pratiche dei dipendenti in continua evoluzione sarà persino maggiore del previsto.** Anche se, indubbiamente, molte funzioni IT meccaniche si sposteranno drasticamente dai lavoratori umani all'intelligenza artificiale, spostare l'AI all'esterno del data center potrà creare opportunità entusiasmanti, persino motivanti, sia per il reparto IT che per i lavoratori aziendali.
4. **Non c'è tempo da perdere.** Le iniziative correlate all'AI sono schizzate in cima alla lista delle priorità strategiche di ogni organizzazione. Se la vostra organizzazione non ha sfruttato completamente l'AI on-premise per provare i nuovi casi di utilizzo e se desiderate saperne di più su cosa può fare l'AI, siete sicuramente indietro rispetto ai tempi. Tuttavia, spostare l'AI vicina agli utenti e nel cloud distribuito rappresenta una nuova possibilità.

Un altro motivo per cui riteniamo che troverete utile questo eBook: alla fine, troverete utili consigli su come la vostra organizzazione possa trarre pieno vantaggio dal passaggio dell'AI nel cloud distribuito.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

Capitolo 1

La rivoluzione dell'AI e come ha già cambiato le dinamiche del mercato

Poiché l'AI è già in circolazione da decenni, è giusto definire l'epoca attuale come una "rivoluzione dell'AI" o l'AI si sta semplicemente evolvendo (anche se ad un ritmo accelerato) dalle sue prime forme e funzioni? Non si tratta di domande retoriche: il nostro settore tende a parlare di rivoluzioni tecniche che irrompono sulla scena apparentemente da un giorno all'altro, essenzialmente inosservate da tutti, tranne dagli esperti che le hanno già sperimentate nei loro laboratori.

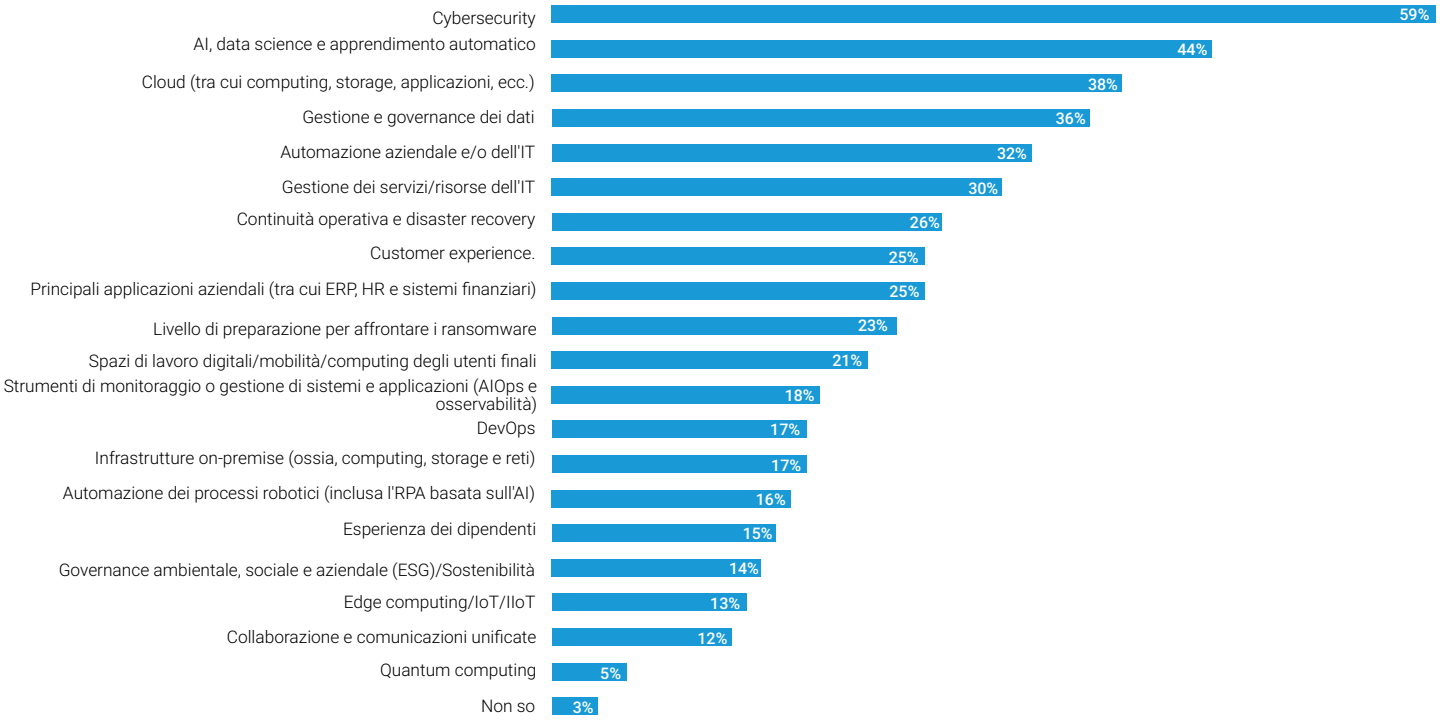
Tuttavia, è giusto definire questo fenomeno una rivoluzione dell'AI per vari motivi. Innanzitutto, è innegabile che la crescita del mercato dell'AI (e l'importanza attribuita dai responsabili aziendali a questa tecnologia) sia un fenomeno straordinario. I suoi tassi di crescita e l'attuale impatto sul mercato superano di gran lunga i movimenti tecnologici a lungo considerati pietre miliari nello sviluppo del settore IT, come i PC, le reti locali, il cloud computing e l'IoT.

I dati riportati di seguito illustrano l'impatto, l'importanza e il potenziale della rivoluzione dell'AI:

- La spesa globale in hardware, software e servizi correlati all'AI ha superato i 600 miliardi di dollari alla fine del 2024 e aumenterà fino a raggiungere i 2,7 trilioni di dollari entro il 2032 con un tasso di crescita annuale composito di oltre il 20%.[1]
- Il 97% dei dirigenti aziendali ha affermato che le loro organizzazioni stanno registrando ritorni positivi sugli investimenti condotti nell'AI.[2]
- La percentuale di CEO che ha inserito l'AI al primo posto nelle loro due principali priorità tecniche è passata dal 4% al 24% in un solo anno.[3]
- Come mostrato nella figura riportata di seguito, negli ultimi due anni, quasi la metà delle aziende statunitensi ha dichiarato che l'AI è molto più importante per il futuro delle loro organizzazioni, superata solo dalla cybersecurity in cima alle iniziative aziendali.[4]

Figura 1. La sicurezza mette l'AI, il cloud e la gestione dei dati al primo posto tra le iniziative tecnologiche cruciali

Quale delle seguenti iniziative tecnologiche sono diventate sempre più importanti per il futuro della vostra organizzazione negli ultimi due anni? (Percentuale di intervistati, N=1, 351, è possibile scegliere più risposte)



1 Fonte: Fortune Business Insights, "Artificial Intelligence market size ... 2024-2032", dicembre 2024.
2 Fonte: Lizzie McWilliams, "Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks", EY.com, 10 dicembre 2024.
3 Fonte: il sondaggio sui CEO di Gartner del 2024 rivela che l'AI è la massima priorità strategica (LinkedIn), maggio 2024.
4 Fonte: risultati completi del sondaggio condotto dall'Enterprise Strategy Group, *2025 Technology Spending Intentions Survey*, dicembre 2024.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

In effetti, l'impennata dell'adozione dell'AI e degli investimenti di capitale ha persino superato quella del cloud computing, uno dei segmenti del panorama IT più dinamici e a più rapida espansione. Il motivo per cui la crescita dell'AI sta eclissando quella del cloud è piuttosto semplice: mentre il cloud computing rappresenta un modo potente e molto pratico per sviluppare e fornire servizi IT, l'AI sta cambiando radicalmente il nostro modo di lavorare, vivere e interagire in ogni aspetto della nostra vita.

Inoltre, la rapida adozione dell'AI ha messo in luce i limiti delle architetture di cloud computing tradizionali. Ad esempio, i modelli cloud centralizzati faticano a soddisfare le notevoli esigenze in termini di performance dovute alla natura dell'AI ad elevato utilizzo di dati in tempo reale. In sostanza, l'AI ha cambiato radicalmente le aspettative per le architetture cloud quando si tratta di questioni come performance, costi, competenze delle persone, sicurezza e governance dei dati.

Cosa sta guidando la crescita stratosferica del mercato dell'AI e il suo impatto sul mercato? Un fattore importante è l'uso molto più perspicace e sofisticato dei modelli di AI da parte delle organizzazioni che si è osservato lo scorso anno. In seguito al rapido aumento dell'utilità e della facilità d'uso dei modelli di AI e allo straordinario ritorno sugli investimenti riscontrato da molte organizzazioni, sono emersi casi di utilizzo importanti sia per rafforzare l'opinione iniziale su come utilizzare l'AI sia per aprire scuole di pensiero completamente nuove su come trarre vantaggio dai modelli di AI.

Ad esempio, consideriamo i seguenti aspetti:

- **La capacità di previsione** è sempre stata una parte essenziale delle organizzazioni di maggior successo e l'AI ha reso questa funzione più precisa, accurata, tempestiva e pratica. Prevedere le future tendenze con strumenti come i modelli di AI predittivi è anche più avanzato rispetto a dieci anni fa quando le organizzazioni hanno adottato le analisi dei big data perché i modelli di AI di oggi sono potenziati con i dati di cui dispongono le organizzazioni. In tal modo, i dirigenti aziendali, non solo i Data Scientist, possono vagliare enormi volumi di dati in modo più rapido e accurato rispetto ai modelli tradizionali.
- **Il riassunto dei documenti**, un tempo, era di dominio degli impiegati, tuttavia, l'AI generativa (GenAI) ha cambiato le regole del gioco. I documenti più lunghi, compresi quelli pieni di contenuti complessi e altamente tecnici, vengono rivisti, analizzati e riassunti automaticamente in tempi ridotti, il che non solo migliora le informazioni fornite, ma consente anche di utilizzare sia i professionisti IT che quelli aziendali in modi più innovativi e strategici, aumentando la produttività e ottimizzando l'experience dei dipendenti.
- **I modelli sui tassi di abbandono** sono ora un elemento critico nella gestione delle relazioni con i clienti perché aiutano le organizzazioni a prendere decisioni migliori più rilevanti da un punto di vista contestuale per prevedere come, quando e perché i clienti potrebbero passare alla concorrenza o perché potrebbero contribuire ad incrementare le vendite.
- **I chatbot basati sull'AI** trasformano il servizio clienti, la gestione dell'assistenza IT e le funzioni delle risorse umane, offrendo servizi automatizzati, intelligenti e in tempo reale per molte richieste di routine. Questi chatbot vengono, persino, utilizzati per applicazioni di ingegneria e informatica altamente tecniche, come la generazione di codice e la modellazione di nuove funzioni.

Questi casi di utilizzo hanno contribuito a generare molte altre applicazioni, workflow e casi di utilizzo dell'AI, come il rilevamento delle minacce alla cybersecurity, l'analisi dei dati/business intelligence, l'analisi della concorrenza, la prototipazione rapida e la gestione della conformità.

Ovviamente, nessuna nuova tecnologia (anche con lo stesso slancio e le stesse promesse dell'AI) è priva di sfide in termini di sviluppo, implementazione e gestione, che possono essere di tipo tecnico, come la mancanza di qualità dei dati per creare la giusta infrastruttura o garantire le elevate performance necessarie per supportare i modelli LLM (Large Language Model), oppure correlate alle aziende o ai rischi, come la governance dei dati per garantire un uso responsabile dell'AI o superare le notevoli lacune di competenze sull'AI sempre crescenti

Tuttavia, dalle indicazioni iniziali emerge come le organizzazioni si rendono conto e stanno affrontando queste sfide con le giuste risorse e il supporto dei dirigenti e dei membri del consiglio di amministrazione. Nel prossimo capitolo, descriveremo uno degli sviluppi chiave che rendono l'AI più di un progetto scientifico: spostare l'attenzione dell'AI dall'addestramento dei modelli all'inferencing.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

Capitolo 2

La svolta rivoluzionaria dell'AI: dall'addestramento all'inferencing

L'addestramento dei modelli di AI è stato un fattore critico nello sviluppo del mercato e, necessariamente, l'aspetto su cui si sono focalizzate inizialmente le organizzazioni nei progetti relativi all'AI. Le organizzazioni hanno capito subito che la mancanza di qualità dei dati (o, forse, la mancanza del tipo e del volume di dati più appropriati) influiscono sullo sviluppo dei modelli di AI. Una ricerca condotta dall'Enterprise Strategy Group di Informa TechTarget ha rivelato che il 37% delle organizzazioni ha citato i problemi legati alla qualità dei dati come la principale sfida che devono affrontare nell'implementazione della GenAI, seguita, al secondo posto, dalle lacune delle competenze sull'AI.[5]

Poiché le organizzazioni si sono rese conto che attingere ai propri dati era il modo migliore per affrontare i problemi legati alla qualità dei dati e per creare modelli LLM più accurati, contestualmente consapevoli e reattivi, ora possono realizzare facilmente il giusto modello di AI. Questo approccio è spesso molto migliore rispetto all'acquisto di modelli di AI commerciali pronti all'uso, che sono stati addestrati su dati di terze parti e, quindi, possono creare errori e imprecisioni dei dati o risultare privi di funzionalità di personalizzazione. Di norma, la maggior parte delle organizzazioni ha focalizzato l'addestramento dei propri modelli LLM in ambienti cloud centralizzati, un approccio efficace per la fase iniziale dello sviluppo di questi modelli.

L'addestramento dei modelli di AI è maturato e si è stabilizzato in termini di qualità e integrità. Le organizzazioni stanno già allocando meno risorse ai propri modelli di addestramento interni, soprattutto perché la regolazione e l'ottimizzazione dell'AI riducono la necessità di disporre di un'infrastruttura incentrata su GPU ad elevato utilizzo di calcolo.

Mike Leone, direttore dell'Enterprise Strategy Group per l'AI, le analisi e la business intelligence, ha condiviso la sua opinione indipendente sulla transizione dall'addestramento all'inferencing, affermando che

"poiché i modelli prestabiliti e le applicazioni basate sull'AI in tempo reale diventano più maturi e prevalenti, l'inferencing svolge ora un ruolo centrale nell'offrire il valore promesso dall'AI in modo più efficiente ed efficace in termini di costi. Questo spostamento di attenzione consente di eseguire implementazioni scalabili in tutti i settori, che vengono favorite, principalmente, dal continuo progresso degli strumenti hardware e software in uso. I vendor sono estremamente focalizzati sulla riorganizzazione e sull'ottimizzazione dell'infrastruttura AI, sul miglioramento dell'efficienza dei modelli e sulla gestione dei costi operativi correnti. Ovviamente, poiché l'inferencing rende l'AI disponibile per tutti, introduce anche sfide come la scalabilità, i problemi di privacy e il controllo normativo".

Il passaggio dall'addestramento dei modelli di AI alle applicazioni basate sull'AI inferencing e i relativi casi di utilizzo ad alto impatto è stato già intrapreso da tempo. Armand Ruiz, vicepresidente di IBM per le piattaforme basate sull'AI, si è espresso brevemente e in modo convincente in questi termini: "Secondo me, l'inferencing dominerà i carichi di lavoro dell'AI. … Una volta addestrato correttamente un modello, idealmente non sarà più necessario ripetere ulteriormente l'addestramento, tuttavia, il processo dovrà avvenire continuamente".[6]

Le transizioni di mercato, quando si verificano, sono, inevitabilmente, caratterizzate da oscillazioni sproporzionate relativamente al modo e alle aree in cui viene investito il capitale. Ad esempio, è stato stimato che circa il 15% della spesa per i chip AI è dedicato all'addestramento dell'intelligenza artificiale, mentre il 45% è destinato all'AI inferencing basato sui data center e il restante 40% all'inferencing basato sull'edge.[7]

5 Fonte: risultati completi del sondaggio condotto dall'Enterprise Strategy Group, "*The State of the Generative AI Market: Widespread Transformation Continues*," settembre 2024.

6 Fonte: LinkedIn, Armand Ruiz, dicembre 2024.

7 Fonte: Jonathan Goldberg, "The AI chip market landscape: Choose your battles carefully", TechSpot.com, 30 maggio 2023.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Da altre ricerche, è emersa una tendenza simile verso l'inferencing, mentre uno studio di Google ha messo in evidenza questa tendenza in un altro modo: Google ha mostrato che il 60% del suo consumo energetico incentrato sull'apprendimento automatico viene utilizzato per l'inferencing rispetto al 40% utilizzato per l'addestramento.[8]

Non sorprende che questo passaggio dall'addestramento dei modelli di AI all'AI inferencing influisca in modo significativo sull'infrastruttura e sulle architetture cloud dell'AI. Questa trasformazione è specialmente importante per le organizzazioni che cercano di utilizzare l'architettura e l'ambiente cloud più appropriati per lo sviluppo, l'implementazione e la gestione delle applicazioni basate sull'AI.

I tradizionali modelli cloud centralizzati funzionano ancora bene per molti carichi di lavoro aziendali, ma l'AI pone sotto pressione l'architettura cloud che deve soddisfare le esigenze in termini di performance dell'AI su larga scala. Queste sfide sembreranno, probabilmente, frustranti problemi di latenza con un notevole divario nel tempo che intercorre tra le query e le risposte ricevute.

Anche l'infrastruttura informatica è un problema chiave di cui le organizzazioni devono tenere conto quando passano dall'addestramento dei modelli di AI all'AI inferencing. Le GPU incentrate sull'AI hanno ricevuto riscontri molto positivi per la loro capacità di fornire la potenza necessaria per addestrare i modelli LLM perché offrono elevati livelli di performance informatiche e larghezza di banda della memoria. Al contrario, l'inferencing richiede una diversa concezione dell'infrastruttura, incentrata sulla bassa latenza e su un uso efficiente delle risorse.

Nel prossimo capitolo, esamineremo più da vicino il modo con cui le organizzazioni possono soddisfare le esigenze in termini di performance dell'AI inferencing, avvicinando le relative risorse agli utenti e posizionandole in un'architettura cloud distribuita.

8 Fonte: "The AI boom could use a shocking amount of electricity", Scientific American, 13 ottobre 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Capitolo 3

L'aumento dell'AI inferencing decentralizzato e del cloud distribuito

Come abbiamo spiegato nel capitolo precedente, la tradizionale architettura cloud centralizzata impone sfide e limiti problematici per le organizzazioni che cercano di ottimizzare il loro utilizzo dell'AI nella creazione di applicazioni innovative e trasformative. È necessario considerare e pianificare varie questioni come la latenza, le limitazioni della larghezza di banda della rete, la necessità di garantire elevate performance su larga scala e la complessità della gestione man mano che i requisiti dell'AI passano dall'addestramento dei modelli all'inferencing.

Le moderne applicazioni di AI inferencing di nuova generazione richiedono la capacità di sviluppare, implementare e sfruttare l'inferencing nelle aree in cui risiedono i dati, ossia sempre più vicino agli utenti o all'interno di un'architettura cloud distribuita, che risulti agile, scalabile e flessibile.

Prima di approfondire l'argomento, dobbiamo chiarire alcune definizioni:

- *AI inferencing decentralizzato*: si tratta di un paradigma per la creazione di workflow dell'AI che includono un data center centralizzato (il cloud) e dispositivi esterni al cloud più vicini agli utenti e alle cose fisiche (l'edge). Questa definizione si contrappone alla pratica più comune in cui le applicazioni basate sull'AI vengono sviluppate ed eseguite interamente nel cloud: ecco perché si è iniziato a parlare di *AI per il cloud*. Inoltre, questa definizione differisce dai precedenti approcci allo sviluppo dell'AI in cui i relativi algoritmi venivano creati sui computer desktop e poi implementati su computer desktop o hardware specializzati per alcune attività, come la lettura dei codici di sicurezza.

La distribuzione dei servizi cloud può aiutare le organizzazioni a conseguire gli obiettivi prefissati in termini di performance delle applicazioni e tempi di risposta, Edge Computing, obblighi normativi o altre esigenze.

- Akamai, un provider di piattaforme per il cloud, la sicurezza e la delivery dei contenuti leader del settore, definisce il *cloud distribuito* come "una forma di cloud computing in cui un'azienda utilizza un'infrastruttura di cloud pubblico in varie località geografiche, mentre le operazioni, la governance e gli aggiornamenti sono gestiti a livello centralizzato da un unico provider di servizi di cloud pubblico. La distribuzione dei servizi cloud può aiutare le organizzazioni a raggiungere gli obiettivi in termini di performance delle applicazioni e tempi di risposta, Edge Computing, obblighi normativi o altre richieste che possono essere soddisfatte fornendo servizi cloud da posizioni più vicine agli utenti finali e ai dispositivi. L'uso del cloud computing distribuito è dovuto all'ascesa delle reti IoT (Internet of Things), dell'intelligenza artificiale e di altri casi di utilizzo che devono elaborare grandi quantità di dati in tempo reale".

Sia l'AI inferencing decentralizzato che il cloud distribuito sono vitali per lo sviluppo e l'utilizzo dell'AI inferencing perché si tratta di metodi moderni, efficienti, sicuri e affidabili per superare le sfide menzionate in precedenza (performance elevate/su larga scala, un uso efficiente delle risorse di elaborazione, storage e rete), nonché per risolvere i problemi di latenza associati ai grandi dataset che, spesso, comprendono enormi quantità di dati non strutturati.

Come avviene tutto ciò nella realtà? Un'azienda, nell'intento di rendere la personalizzazione un tratto distintivo per le sue customer experience, aveva bisogno di acquisire i dati da un'ampia gamma di dispositivi e fonti di dati, pertanto aveva deciso di utilizzare un chatbot AI basato sui modelli LLM per distribuire i dati localizzati vicino agli utenti. In tal modo, ha ridotto la latenza e migliorato i tempi di risposta, incrementando così notevolmente le performance e offrendo un supporto ai clienti più rapido e personalizzato.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

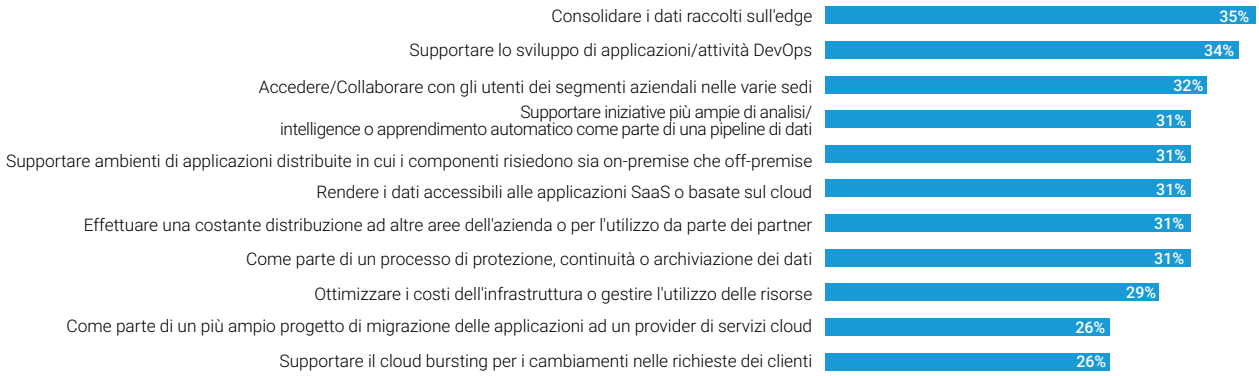
Questo approccio è sempre più tipico delle organizzazioni che cercano di trarre vantaggio dalle loro crescenti opportunità con le applicazioni basate sull'AI inferencing. Per risolvere i problemi di performance influenzati dalle limitazioni delle infrastrutture informatiche e dai problemi di latenza, i carichi di lavoro dell'AI si spostano più vicino alle fonti dei dati utilizzate da un maggior numero di utenti tramite l'AI inferencing decentralizzato e le architetture del cloud distribuito.

Ad esempio, questo modello è ampiamente usato in settori caratterizzati da sedi altamente distribuite e dislocate in varie aree geografiche, tra cui il retail, le telecomunicazioni e i media. Questo approccio sta diventando sempre più il modello dell'architettura standard per l'AI inferencing grazie alla sua capacità di superare i problemi di latenza e performance su larga scala associati ai complessi requisiti dell'AI inferencing.

Un aspetto importante che ha favorito il passaggio ad un modello di cloud distribuito per l'AI inferencing è rappresentato dalla crescente esigenza avvertita da parte delle organizzazioni di utilizzare l'edge per spostare i dati tra i data center e i provider di servizi cloud (CSP). L'Enterprise Strategy Group ha rivelato che il 35% delle organizzazioni spostano regolarmente i dati tra i data center e l'infrastruttura del proprio CSP per consolidare i dati raccolti sull'edge: è questa la loro principale motivazione alla base dello spostamento dei dati.[9]

Figura 2. I principali fattori sull'edge che hanno favorito lo spostamento dei dati tra i data center e i servizi del cloud pubblico

Poiché avete indicato che le vostre organizzazioni spostano regolarmente i dati tra i propri data center e i servizi del cloud pubblico, potete precisare per quale scopo? (Percentuale di intervistati, N=170, è possibile scegliere più risposte)



9 Fonte: rapporto di ricerca dell'Enterprise Strategy Group, *Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud*, novembre 2023.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

Capitolo 4

Perché l'AI inferencing decentralizzato è il futuro

La creazione e l'implementazione dell'AI inferencing con i dati raccolti sull'edge sono operazioni fondamentali per aiutare le organizzazioni a trarre il massimo vantaggio dall'inferencing. Considerando il notevole incremento dei dati che vengono creati e che sono presenti su vari dispositivi sull'edge, tra cui smartphone, dispositivi palmari appositamente progettati, segnali digitali e dispositivi IoT, le organizzazioni devono adottare una visione dell'edge tale da ottenere l'elaborazione dei dati in tempo reale come richiesto dall'AI inferencing.

La tendenza verso l'AI inferencing decentralizzato per soddisfare i requisiti di questa tecnologia è stata rapida e profonda. Inoltre, la ricerca rivela che la spesa necessaria per avvicinare i sistemi di inferencing agli utenti aumenterà di 10 volte, passando da 27 miliardi di dollari registrati nel 2024 a quasi 270 miliardi di dollari stimati per il 2032.[10]

L'AI inferencing decentralizzato prevede una serie di considerazioni importanti che consentono di aiutare le organizzazioni a sviluppare e distribuire le applicazioni di AI inferencing, come l'elaborazione dell'AI nativa dei dispositivi, i sistemi di sicurezza integrati che sono stati ottimizzati per i dispositivi sull'edge, il basso consumo energetico e la latenza ridotta. Queste funzionalità si traducono in performance più rapide su larga scala, informazioni sui dati più accurate e approfondite e minori investimenti in larghezza di banda, che, a loro volta, consentono ai membri di un'organizzazione dislocati in varie aree geografiche di utilizzare l'AI inferencing per prendere decisioni più intelligenti e in tempo reale, molto più vicine alle aree in cui i dati vengono creati, archiviati, elaborati e resi accessibili.

Bloomfield è un esempio di un'organizzazione lungimirante che ha utilizzato l'AI inferencing decentralizzato per migliorare le operazioni essenziali e per semplificare la gestione dei dati. Dovendo gestire enormi quantità di dati provenienti dalle sue videocamere intelligenti dislocate in vari punti in modo efficiente, sicuro e vantaggioso in termini di costi per gli agricoltori che, spesso, lavorano con un accesso limitato alla rete, Bloomfield ha sfruttato l'AI inferencing con un modello di architettura decentralizzato per stabilire come organizzare, archiviare e visualizzare al meglio i dati per gli agricoltori. Questo sistema ha semplificato la gestione delle colture, ha migliorato i raccolti e ha reso i processi decisionali più rapidi e accurati, anche se ben lontani dai data center on-premise.

La capacità dell'AI inferencing decentralizzato di supportare le applicazioni basate su questa tecnologia, che sono distribuite in varie posizioni, la rende ideale per casi di utilizzo simili ai seguenti:

- **Diagnosi cliniche e telemedicina,** in cui le decisioni vengono prese dai medici durante i loro spostamenti tramite l'utilizzo di smartphone, dispositivi IoT e altri dispositivi portatili leggeri, ma potenti. L'AI inferencing decentralizzato è anche ideale per prendere decisioni in tempo reale tramite dispositivi sanitari intelligenti, come pompe infusionali, pacemaker, cardiofrequenzimetri e strumenti per analisi mediche.
- **Città intelligenti,** in cui i sistemi sull'edge eseguono tutte le operazioni necessarie, dal coordinamento del traffico al monitoraggio dei sistemi di pubblica utilità fino alla semplificazione della registrazione degli elettori e al miglioramento della sicurezza pubblica.
- **Veicoli senza conducente,** in cui vengono utilizzate videocamere intelligenti e i dati dei sensori di bordo per valutare i percorsi disponibili, per monitorare e migliorare il consumo di carburante, per documentare i cambiamenti del GPS e per comunicare con le forze dell'ordine e gli enti di pubblica sicurezza.
- **Nel retail,** che supporta un'ampia gamma di applicazioni basate sull'AI inferencing, dalla videosorveglianza IP e dalla gestione dell'inventario alla prevenzione dei furti fino ai modelli del traffico dei clienti in-store.

Mentre le aziende cercano il modo di ridurre la latenza nelle risposte alle query in tempo reale, l'AI inferencing decentralizzato diventerà un requisito critico per prendere decisioni aziendali più efficienti, informate e accurate. Le applicazioni basate sull'AI non dovranno più inviare o richiedere dati a data center remoti, ma potranno elaborare e condividere i dati su un cloud distribuito a livello globale.

Il direttore della progettazione di un'importante società di tecnologie pubblicitarie ha affermato: "Abbiamo impiegato tantissimo tempo per ottimizzare il nostro modello di AI. Ora, ci siamo resi conto che l'inferencing non può passare in secondo piano".

Le organizzazioni non possono più concedersi il lusso di decidere se adottare o meno l'AI inferencing come principale componente della propria strategia aziendale, ma devono utilizzare l'AI inferencing decentralizzato con la massima urgenza perché molti dei loro concorrenti hanno già compiuto importanti passi in avanti verso l'utilizzo di questa tecnologia vicino agli utenti.

10 Fonte: Fortune Business Insights, "Edge AI Market Size … 2024-2032", dicembre 2024.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Capitolo 5

Uno sguardo strategico all'AI, all'edge e al cloud distribuito: il loro futuro e le relative implicazioni

La valutazione, la scelta e la collaborazione con un partner tecnologico affidabile sono componenti fondamentali per il successo di un progetto di AI inferencing di un'organizzazione, che, tuttavia, presenta varie difficoltà.

Si tratta di aspetti fondamentali per un fatto innegabile: il sostanziale divario di competenze nell'AI quando si tratta di infrastruttura, architettura nel cloud, casi di utilizzo dell'inferencing, ottimizzazione delle applicazioni e implementazione impeccabile,

ma sono anche complicati a causa della mancanza di candidati adatti. Come abbiamo notato in precedenza, per la riuscita di un progetto di AI inferencing è necessario combinare perfettamente know-how tecnico, competenze del settore, conoscenze dei sistemi di sicurezza, governance e gestione dei rischi, oltre ad un'attenzione maniacale per i dettagli nell'implementazione e nella gestione continua.

Per sviluppare e implementare l'AI inferencing in tutti gli spazi fisici e virtuali di un'azienda (dal core all'edge fino al cloud e viceversa), le organizzazioni devono rivolgersi ad un provider con una precedente esperienza nel settore. Disporre di un partner in grado di comprendere sia gli elementi tecnici che quelli commerciali di un progetto di AI è un elemento imprescindibile per garantirne il successo.

Akamai, un importante fornitore di avanzate soluzioni per il cloud computing, la sicurezza e la delivery dei contenuti, vanta una solida esperienza nella creazione, nell'implementazione e nella gestione di iniziative basate sull'AI, dall'addestramento dei modelli all'inferencing. La sua esperienza pratica con carichi di lavoro complessi ad elevato utilizzo di calcolo consente all'azienda di collaborare con un'ampia fascia di clienti in diversi casi di utilizzo, settori e aree geografiche.

Per aiutare le organizzazioni a soddisfare facilmente i loro requisiti di AI inferencing, Akamai ha implementato con successo, in modo coerente, affidabile e sicuro, un'architettura cloud distribuita per consentire alle organizzazioni di elaborare i carichi di lavoro dell'AI più vicino agli utenti finali. Questo approccio si è rivelato prezioso per ridurre la latenza e garantire elevate performance su larga scala, due aspetti ugualmente fondamentali per i carichi di lavoro dell'AI inferencing ad elevato utilizzo di calcolo.

Uno dei vantaggi principali dell'architettura cloud distribuita di Akamai consiste nella capacità degli sviluppatori di focalizzarsi esclusivamente su uno sviluppo indipendente dalla piattaforma in un sistema cloud-native basato su standard aperti per il cloud.

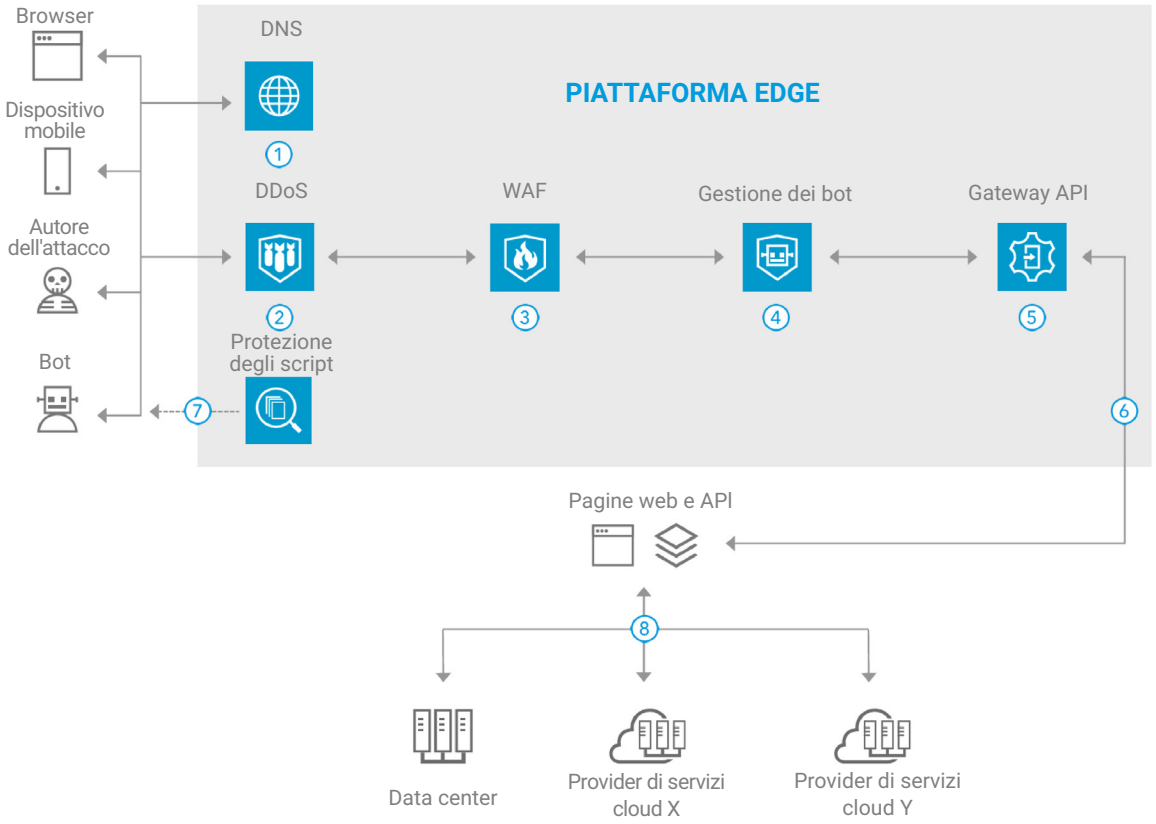
Per offrire elevate performance su larga scala, Akamai utilizza varie tecniche di caching, integrate con il software open source, per garantire un throughput adeguato e per limitare la latenza. Ad esempio, molti clienti di Akamai che utilizzano la sua architettura cloud distribuita hanno riferito di aver raggiunto una riduzione del 50% nei tempi di risposta per i chatbot progettati per i clienti, un caso di utilizzo dell'inferencing molto importante e popolare. Questo tipo di risposta quasi in tempo reale consente alle organizzazioni che operano in un'ampia gamma di settori, tra cui il retail, il settore sanitario, i servizi finanziari e l'high-tech, di fornire un servizio clienti più rapido e personalizzato in base alle loro specifiche esigenze.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Akamai, inoltre, offre un'architettura di riferimento basata sull'AI (vedere la Figura 3 di seguito), che consente alle organizzazioni di essere subito operative e di raggiungere un elevato valore economico e produttivo. Questa architettura di riferimento è completa in quanto, tramite una serie consueta di browser e dispositivi definiti dall'utente, le organizzazioni possono implementare una soluzione aziendale basata sull'AI dal data center principale all'edge fino al cloud e viceversa. In tal modo, possono trarre vantaggio tutte le parti coinvolte nell'azienda: sviluppatori, esperti IT, responsabili dell'infrastruttura, architetti del cloud e proprietari di segmenti aziendali.

Figura 3.

Considerando la rapidità delle innovazioni tecnologiche e l'aumento della complessità e del valore dei carichi di lavoro dell'AI, le organizzazioni devono procedere velocemente e in modo deciso per superare la concorrenza. È fondamentale sfruttare appieno il potenziale offerto dall'AI inferencing per adottare l'AI inferencing decentralizzato e un'architettura cloud distribuita, soprattutto se il processo viene pianificato, sviluppato, implementato e gestito da un'azienda leader di fama consolidata nel settore come Akamai.



```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Conclusione

Cosa deve fare la vostra organizzazione

Abbiamo già detto che l'AI ha cambiato le regole del gioco per realizzare l'eccellenza dell'IT e la trasformazione aziendale. Tuttavia, la chiave per ottimizzare il valore degli investimenti nell'AI di un'organizzazione è procedere rapidamente dall'addestramento dei modelli di AI all'AI inferencing.

Per trarre il massimo vantaggio da tutto ciò che l'AI inferencing ha da offrire, le organizzazioni devono sviluppare e implementare le loro applicazioni e i carichi di lavoro basati sull'inferencing più vicino alle aree in cui viene svolto il lavoro e in cui i loro dipendenti più importanti possono svolgere egregiamente il loro lavoro con i sistemi in uso. Pertanto, è necessario passare da un modello di cloud computing centralizzato ad un modello di cloud distribuito, che è plasmato notevolmente dall'AI sull'edge.

Come abbiamo sottolineato in precedenza in questo eBook, le organizzazioni devono considerare molti problemi e devono comprendere e superare numerose sfide, il che può sembrare scoraggiante se non hanno ancora assegnato una priorità strategica ai progetti relativi all'AI, ma si tratta di passi importanti che possono e devono intraprendere per progredire e per tenersi al passo con la concorrenza.

"Le organizzazioni che cercano di mettersi in pari con l'AI non devono farsi prendere dalla paura di perdere l'opportunità di tenere il passo con la tecnologia", ha sottolineato Leone dell'Enterprise Strategy Group. "È necessario focalizzarsi sulla costruzione delle giuste fondamenta adottando misure come l'identificazione di specifici problemi aziendali che l'AI può risolvere, sulla valutazione dell'infrastruttura di dati esistente e delle competenze interne e sull'identificazione dei partner chiave in grado di colmare le eventuali lacune. Si tratta di mosse strategiche in quanto identificano uno o due progetti pilota ad alto impatto e casi di utilizzo in grado di dimostrare un chiaro valore tramite i dati esistenti di elevata qualità.

Sapere cosa misurare e come misurarlo è fondamentale per il successo di qualsiasi iniziativa di AI inferencing.

Nello stesso tempo, inoltre, è necessario sviluppare talenti interni e stabilire quadri di governance per garantire un uso responsabile. L'importante è evitare un'implementazione affrettata favorendo l'esecuzione di operazioni strategiche in grado di costruire funzionalità fondamentali e, idealmente, durature".

A questo scopo, ecco quattro passaggi di cui le organizzazioni dovrebbero tener conto e mettere in pratica il prima possibile:

1. **Fare attenzione a scegliere bene uno specifico caso di utilizzo.** Si tratta di un primo passaggio essenziale perché le organizzazioni devono avere fiducia nella propria capacità di riuscire ad implementare correttamente l'AI inferencing. Inoltre, è importante creare fiducia tra le parti interessate all'interno dell'azienda e i responsabili dell'approvazione finanziaria con risultati concreti fin da subito.
2. **La scelta dell'infrastruttura fa la differenza.** Ovviamente, poiché l'AI inferencing è una tecnologia ad elevato utilizzo di calcolo, la decisione più semplice consiste nel scegliere la GPU leader del settore come componente fondamentale del sistema in uso. Tuttavia, esistono molte eccellenti GPU e altre infrastrutture dell'AI tra cui scegliere. Occorre prendersi il tempo necessario per fare la scelta giusta (o, meglio, affidarsi alle conoscenze e alle competenze di un partner con una precedente esperienza nel settore).
3. **Accaparrarsi il sostegno dei dirigenti.** In questo caso, bisogna focalizzarsi sulle due principali preoccupazioni di ogni dirigente e membro del consiglio di amministrazione: quali sono le nostre opportunità e come si bilanciano con i potenziali rischi? Il progetto e i casi di utilizzo in questione devono allinearsi con i principali obiettivi aziendali, come migliorare le customer experience, identificare e sfruttare le opportunità del mercato più rapidamente e creare un processo di test beta più accurato, identificando e assicurandosi, al contempo, sponsor esecutivi appartenenti a vari gruppi, come l'IT, i segmenti aziendali e i membri del consiglio di amministrazione.
4. **Definire e realizzare il successo.** Sapere cosa misurare e come misurarlo è fondamentale per il successo di qualsiasi iniziativa di AI inferencing. A tal scopo, è necessario scegliere metriche (o collaborare con il proprio partner per sviluppare criteri personalizzati) realistiche, rilevanti e sufficientemente solide da poter conseguire facilmente risultati positivi, specialmente dopo aver osservato i vantaggi apportati con i primi casi di utilizzo dell'AI inferencing.

Questa ricerca è stata commissionata da Akamai e prodotta da TechTarget Inc.