



Autore: MIKE HILTON
Vicepresidente, North America
Hybrid Solutions, HPE

Il fattore segreto che determina il successo dell'AI: il cloud ibrido

Vuoi superare gli ostacoli più comuni che si frappongono ai tuoi progetti di AI? Ecco cinque aspetti chiave in cui la tecnologia ibrida può essere determinante.

L'INTELLIGENZA ARTIFICIALE (AI) sta dilagando in tutti i settori a un ritmo incalzante, promuovendo importanti innovazioni e potenziando l'efficienza. Ma inserire l'AI nelle operazioni di business quotidiane non è semplice. Comporta una serie di problematiche, tra cui quelle legate alla sicurezza dei dati, alla scalabilità, alla gestione dei costi e alla sostenibilità.

È qui che entrano in gioco le piattaforme di cloud ibrido con la loro combinazione equilibrata tra le ingenti risorse del cloud pubblico e la sicurezza e il controllo dell'infrastruttura on-premise. Si tratta di una soluzione ideale per la gestione dei carichi di lavoro AI.

Il vantaggio di un approccio ibrido è che non richiede il trasferimento di tutte le operazioni in un nuovo data center o l'adozione esclusiva di un cloud pubblico hyperscale. Questo tipo di flessibilità è essenziale per le aziende che hanno bisogno di spazio per la scalabilità e non possono permettersi di scendere a compromessi nella gestione delle operazioni

o nella sicurezza dei dati. Ecco cinque aspetti essenziali del cloud ibrido che possono aiutarti a ottenere il massimo dalle tue iniziative AI.

1. Dati disponibili ovunque

I dati sono la linfa vitale dell'AI, ma selezionarli, gestirli e metterli in sicurezza su vasta scala comporta problematiche rilevanti. L'addestramento dei modelli di AI non è un'attività una tantum; gli algoritmi devono essere continuamente perfezionati per migliorare la precisione, adattarsi a nuovi dati e ottimizzare le prestazioni. Questo processo costante richiede la prossimità dei dati al carico di lavoro per facilitare l'elaborazione in tempo reale e l'apprendimento continuo.

Purtroppo, in molte organizzazioni i dati sono dislocati in varie posizioni, cloud e on-premise. Trasferire petabyte di dati su reti distribuite o tra ambienti cloud non è né efficiente né economicamente vantaggioso.

Una piattaforma di cloud ibrido offre una soluzione pratica. Non è necessario consolidare tutti i tuoi dati in un unico data center o passare completamente a un cloud pubblico hyperscale. Puoi invece mantenere i tuoi dati all'edge e integrarli alla perfezione in un'architettura di cloud ibrido. Questa configurazione consente al modello AI di accedere e utilizzare efficacemente i dati non appena entrano nella fase di produzione, rendendoli immediatamente disponibili ai team che ne hanno bisogno.

2. Sicurezza e compliance protette

Quando si distribuisce l'AI, è fondamentale mettere in sicurezza e gestire i dati. La natura riservata dei dati utilizzati nelle applicazioni di AI li rende un obiettivo privilegiato

Con il cloud ibrido, i dati possono rimanere in un data center all'interno della giurisdizione sovrana di tua scelta, pur contribuendo all'addestramento e alle applicazioni di AI.

delle violazioni della sicurezza, con conseguenti gravi ripercussioni. Nessuna organizzazione vuole rischiare di divulgare i propri dati aziendali o la proprietà intellettuale affidandosi esclusivamente a un cloud pubblico.

I modelli di cloud ibrido affrontano direttamente queste problematiche di sicurezza, consentendo la conservazione dei dati critici on-premise in base a rigorosi protocolli di sicurezza e l'impiego della potenza di elaborazione del cloud per elaborare le informazioni meno sensibili. Questo significa che i dati possono rimanere in un data center all'interno della giurisdizione sovrana di tua scelta, pur contribuendo all'addestramento e alle applicazioni di AI. Questi obiettivi vengono raggiunti grazie a una connessione operativa fluida che si estende sia agli ambienti cloud che ai data center on-premise.

I modelli di cloud ibrido sono particolarmente vantaggiosi per le organizzazioni che devono rispettare severe normative sui dati in diverse ubicazioni. Facilitando l'elaborazione e lo storage dei dati on-premise, in cloud privati o in cloud pubblici situati in regioni diverse, i cloud ibridi consentono di ottemperare alle leggi sulla sovranità e sull'ubicazione dei dati.

Questa flessibilità aiuta le organizzazioni a soddisfare i requisiti normativi fondamentali, garantendo l'agilità necessaria per acquisire un vantaggio competitivo.

3. Scalabilità verticale bidirezionale

La scalabilità può rappresentare una grande problematica quando si eseguono carichi di lavoro AI nel cloud pubblico. I carichi di lavoro AI possono essere molto variabili e richiedono una scalabilità verticale rapida in base all'intensità del carico di lavoro, ad esempio durante i periodi di addestramento dei modelli di machine learning. Sebbene le risorse del cloud possano essere modulate dinamicamente, il processo effettivo può risultare complesso e non sempre immediato. Eventuali ritardi nell'adeguamento della scala possono portare a colli di bottiglia nelle prestazioni o a costi inutili.

L'integrazione di un modello di cloud ibrido con la containerizzazione offre una soluzione efficace per gestire tali carichi di lavoro AI. La containerizzazione, un concetto chiave nello sviluppo software, semplifica la distribuzione, la scalabilità e la gestione delle applicazioni incapsulandole in container. Tali container sono unità leggere ed eseguibili che racchiudono il codice e tutte le sue dipendenze, garantendo

l'esecuzione omogenea ed efficiente dell'applicazione in ambienti di elaborazione diversi.

Questo approccio si rivela particolarmente vantaggioso per controllare i costi e ridurre al minimo il consumo di energia. Con la containerizzazione all'interno di un modello di cloud ibrido, puoi sviluppare le tue soluzioni di AI nel data center, in un ambiente di colocation o all'edge della tua infrastruttura e distribuirle quando necessario. Per i casi in cui hai bisogno di una scalabilità rapida e consistente, ad esempio un aumento di 200 o 300 volte durante i periodi di picco, puoi estendere la tua capacità al cloud pubblico per sfruttarne l'elasticità. In seguito, potrai ridimensionarti con la diminuzione della domanda, ottimizzando sia le prestazioni che l'efficacia dei costi.

4. Riduzione dell'impatto su Madre Terra

L'impatto dell'AI sull'ambiente non è trascurabile. Le problematiche legate alla sostenibilità derivano principalmente dal tributo ambientale rappresentato dalle risorse di elaborazione necessarie per le esigenze di addestramento e di funzionamento dell'AI. Questo comprende l'uso di GPU e

¹"Power of AI: Wild predictions of power demand from AI put industry on edge", S and P Global Commodity Insights, 16 ottobre 2023.

CPU, infrastruttura di rete, operazioni di data center e sistemi di storage, tutti elementi che richiedono un'alimentazione costante. Considerata l'attuale traiettoria di sviluppo dell'AI, si prevede che i carichi di lavoro AI potrebbero rappresentare fino al 4% del consumo energetico mondiale entro il 2030.¹

I modelli di cloud ibrido attenuano molti di questi problemi di sostenibilità ottimizzando l'allocazione e l'utilizzo delle risorse di elaborazione, con una significativa riduzione del consumo energetico e dei costi. Ad esempio, i carichi di lavoro possono essere gestiti con data center privati ad alta efficienza energetica per le attività meno impegnative e attingendo a risorse cloud pubbliche scalabili per le attività di elaborazione più impegnative, a seconda delle necessità. Tale strategia non solo contribuisce alla riduzione dei costi di mantenimento dei data center sottoutilizzati, ma allinea anche il consumo energetico alle effettive esigenze di elaborazione.

Inoltre, molti provider di colocation stanno investendo in fonti di energia rinnovabile e in tecnologie di raffreddamento innovative che superano di gran lunga l'offerta dei data center privati. Tale investimento consente alle organizzazioni di creare cloud ibridi in questi ambienti e distribuire i carichi di lavoro AI senza dover effettuare upgrade diretti della sostenibilità nei propri data center.

Il consumo di energia può essere ulteriormente ridotto affinando i modelli AI. Aniché riaddestrare un intero modello, potrebbe essere necessario mettere a punto solo alcuni elementi chiave, riducendo in modo significativo il consumo di energia nel tempo. Inoltre, l'utilizzo di modelli AI preconfezionati che richiedono l'addestramento solo su dati specifici dell'organizzazione può rendere le implementazioni di AI molto più efficienti dal punto di vista energetico. Sfruttando questi modelli preesistenti, le aziende possono incrementare sensibilmente l'efficienza energetica delle operazioni AI.

5. Un reparto contabilità soddisfatto

L'esecuzione di carichi di lavoro AI nel cloud pubblico può comportare costi significativi. I cloud pubblici fanno pagare l'uso di CPU e GPU, lo spostamento dei dati e lo storage, il che può far lievitare notevolmente le spese, soprattutto

per le attività di AI a uso intensivo di dati. Inoltre, la natura scalabile dei cloud pubblici, pur risultando vantaggiosa, può comportare spese imprevedibilmente elevate durante i picchi di attività. Il costo dell'addestramento di un singolo Large Language Model, che può durare settimane o mesi, è un esempio del potenziale onere finanziario dei carichi di lavoro AI basati su cloud. I modelli di cloud ibrido rappresentano un metodo economicamente vantaggioso per gestire le spese. Puoi ottimizzare l'utilizzo delle risorse mantenendo le operazioni di routine e lo storage dati on-premise o in un cloud privato, dove i costi sono in genere più bassi e passando, se necessario, ai cloud pubblici solo per le attività a uso intensivo di elaborazione. Questo approccio consente di ridurre al minimo i trasferimenti di dati non necessari, riducendo i costosi canoni della larghezza di banda e migliorando il controllo dei costi per lo storage dati.

Il cloud ibrido può inoltre offrire una maggiore flessibilità

Con il cloud ibrido, puoi realizzare una funzionalità core che gestisce i costi operativi con un controllo e una prevedibilità migliori, ma che all'occorrenza può scalare in orizzontale verso un servizio di cloud pubblico nelle ore non di punta, contribuendo a evitare l'overprovisioning.

nelle modalità di fornitura di infrastrutture ad alte prestazioni per container e AI che utilizzano le GPU. Con il cloud ibrido, puoi creare una funzionalità core che gestisce i costi operativi con un controllo e una prevedibilità migliori, ma che all'occorrenza può scalare in orizzontale verso un servizio di cloud pubblico nelle ore non di punta, contribuendo a evitare l'overprovisioning. In definitiva, i modelli di cloud ibrido offrono una soluzione efficiente per bilanciare i requisiti in termini di prestazioni con i vincoli di budget, consentendoti di sfruttare efficacemente la potenza dell'AI e di migliorare al contempo l'efficienza dei costi.

Fai lavorare l'AI in modo più intelligente con il cloud ibrido

Sfruttare la potenza delle architetture di cloud ibrido è un passo fondamentale per gestire le problematiche dei carichi di lavoro AI. Ma l'adozione efficace di un modello di cloud ibrido non riguarda solo la tecnologia in sé, bensì la comprensione e l'allineamento della tecnologia con specifici obiettivi di business. Molte iniziative falliscono non per inadeguatezza tecnica, ma per mancanza di un chiaro allineamento con risultati di business tangibili. Ad esempio, l'integrazione dell'AI per migliorare le capacità dei team del servizio clienti attraverso modelli avanzati di elaborazione del linguaggio naturale può trasformare decenni di dati dell'help desk in informazioni fruibili, migliorando sostanzialmente la qualità del servizio.

Questo tipo di applicazioni illustra il potenziale dell'AI nel fornire risultati significativi quando viene correttamente orientata verso obiettivi ben definiti. Un ambiente di cloud ibrido supporta questo aspetto fornendo l'infrastruttura versatile e scalabile necessaria per distribuire soluzioni AI in grado di adattarsi ed evolversi in linea con le esigenze aziendali strategiche. Consente alle organizzazioni di gestire i dati in modo efficace, garantendo governance, riducendo al minimo il bias e fornendo al contempo la potenza di elaborazione necessaria per elaborare rapidamente grandi set di dati.

Se implementato nel modo giusto, il cloud ibrido non è solo una scelta tecnologica, ma un fattore strategico che aiuta le aziende a sfruttare l'AI per ottenere risultati specifici e significativi.

L'autore

Mike Hilton è Vicepresidente della divisione Hybrid Solutions per il Nord America presso HPE. In precedenza, è stato presidente di HPE Canada e, prima di entrare in HPE, Mike ha ricoperto ruoli imprenditoriali e di alta dirigenza, spaziando tra operazioni aziendali, vendite, marketing e servizi tecnici.

[HPE.com/ai/insights](https://www.hpe.com/ai/insights)

THE DOPPLER

Una rivista digitale in cui gli innovatori condividono strategie tecnologiche, informazioni dettagliate e progressi sull'AI e sulla trasformazione dell'IT.

Leggi l'intero numero di
[The Doppler - the AI issue](#)