TheFuturum Group

Liberare il potenziale dell'AI: come sfruttare HPE Private Cloud AI per accelerare la distribuzione dell'AI e promuovere l'innovazione

AUTORE

Steven Dickens
Vice President e Practice Lead | The Futurum Group

IN COLLABORAZIONE CON

Hewlett Packard Enterprise

NVIDIA

GIUGNO 2024

Documento di sintesi

Il potenziale rivoluzionario dell'AI generativa (GenAI) ha suscitato un notevole interesse e stimolato la crescita in diversi settori. Solo nel 2023, il numero dei progetti GenAI pubblici in GitHub ha registrato un'impennata del 248%, sottolineando l'aumento esponenziale dell'adozione delle tecnologie AI. Tuttavia, questa rapida espansione ha portato alla luce una serie di problematiche. Molte aziende presentano un panorama tecnologico molto complesso, che spesso ostacola la transizione dal progetto AI pilota all'implementazione in produzione. Inoltre, per le imprese che temono di esporre i dati proprietari ai modelli pubblici, la protezione dalla proprietà intellettuale assume un'importanza prioritaria. Collaborando con NVIDIA, HPE risolve tutti questi problemi offrendo soluzioni AI di cloud privato, che sono scalabili, sicure e ottimizzate per aumentare le prestazioni al massimo. Queste soluzioni offrono alle aziende l'opportunità di sfruttare il potenziale dell'AI, mantenendo il controllo dei propri dati e della propria infrastruttura, mentre accelerano il percorso dell'innovazione basata sull'AI.

Panoramica della soluzione HPE Private Cloud AI

Parte del portafoglio NVIDIA AI Computing by HPE, HPE Private Cloud AI è un cloud privato chiavi in mano, scalabile e ottimizzato per l'AI, espressamente progettato per accelerare la distribuzione dei progetti AI assicurando al contempo che i dati rimarranno sempre sotto il controllo dell'azienda. La soluzione combina le risorse di elaborazione accelerata, rete e software NVIDIA con lo storage e le risorse di elaborazione ad alte prestazioni di HPE, il cloud HPE GreenLake e le funzionalità HPE per pipeline dei dati, orchestrazione ed MLOps. In questo modo, offre ad aziende di ogni dimensione un percorso rapido e flessibile per lo sviluppo e la distribuzione delle applicazioni GenAI. Sfruttando queste capacità, è possibile superare gli ostacoli all'adozione dell'AI, assicurando un time to value più rapido e una scalabilità trasparente a supporto della crescita futura.

Dichiarazione d'intenti per la sintesi della ricerca

Questa sintesi della ricerca ha lo scopo di fornire un'analisi approfondita delle problematiche che le imprese che distribuiscono i carichi di lavoro AI devono fronteggiare, spiegando come risolverle efficacemente tramite HPE Private Cloud AI. Nella sintesi vengono evidenziati i risultati e i dettagli principali relativi alle capacità della soluzione, in termini di accelerazione del time to value, ripetibilità e scalabilità. Offre inoltre consigli utili alle aziende che desiderano ottimizzare le distribuzioni AI, in modo da sfruttare al massimo i vantaggi delle tecnologie AI e al contempo gestire i costi e garantire la sicurezza.

Risultati e dettagli principali

1. Accelerazione del time to value: la soluzione di HPE e NVIDIA riduce drasticamente il time to market per i progetti AI, offrendo tool, modelli e infrastrutture preintegrati e collaudati. Questo approccio consente alle imprese di passare rapidamente dalla fase pilota all'implementazione in produzione completa, accelerando la realizzazione concreta dei vantaggi dell'AI.

2. Soluzioni ripetibili: l'infrastruttura standardizzata chiavi in mano offerta da HPE Private Cloud AI garantisce distribuzioni AI scalabili, coerenti e ad alte prestazioni. Sfruttando gli strumenti predefiniti, integrati e collaudati, è possibile distribuire i carichi di lavoro AI in modo semplice e rapido, spesso con un solo clic.

3. Esperienza cloud: espressamente progettata per scalare di pari passo con le esigenze aziendali, la soluzione HPE offre opzioni di distribuzione flessibili e modelli di pagamento a consumo, che supportano la crescita permettendo di gestire efficacemente i costi. Il modello CapEx e la gestibilità della soluzione aumentano ulteriormente la scalabilità, permettendo alle aziende di espandere le funzionalità AI in modo trasparente.

Di seguito è riportata una panoramica aggiornata dei nuovi aspetti da considerare nella presentazione di un business case che privilegia il cloud ibrido/privato al cloud pubblico per l'AI, illustrando le problematiche specifiche:

Panoramica delle problematiche aziendali nella distribuzione dell'AI

Rispetto al cloud pubblico, le soluzioni di cloud ibrido/privato si prestano maggiormente alla risoluzione delle problematiche specifiche che le aziende che valutano l'implementazione dell'AI devono fronteggiare:

1. Definizione, sviluppo e operazionalizzazione dei casi d'uso appropriati: è necessario identificare accuratamente i casi d'uso dell'AI che generano il massimo valore. A tale scopo, occorre comprendere a fondo le esigenze aziendali e avere la capacità di tradurle in progetti AI. Una volta identificati i casi d'uso, la fase di sviluppo e operazionalizzazione è molto complessa perché, solo per citare qualche esempio, richiede l'integrazione dei modelli AI nei flussi di lavoro esistenti, in modo da garantire la produzione di risultati aziendali concreti.

2. Massimizzazione dell'impatto sui dati, continuando a proteggere la proprietà intellettuale: è essenziale sfruttare efficacemente i dati nei modelli AI, per generare informazioni approfondite veramente utili, ma questo deve essere fatto senza compromettere la proprietà intellettuale sensibile. Le soluzioni di cloud pubblico suscitano spesso molti dubbi in merito al controllo e alla sicurezza dei dati, poiché complicano la protezione di algoritmi proprietari e dati sensibili dalle violazioni e dagli accessi non autorizzati.

3. Espansione delle iniziative AI evitando l'aumento esponenziale dei costi: scalare le iniziative AI per far fronte a esigenze in continuo aumento può generare costi proibitivi, soprattutto quando tutto dipende dalle risorse di un cloud pubblico. Le aziende devono bilanciare le esigenze di potenza di elaborazione e storage con le implicazioni finanziarie, garantendo distribuzioni AI economicamente vantaggiose anche in caso di espansione.

4. Ottimizzazione della tecnologia per casi d'uso specifici: il problema è che le diverse applicazioni AI richiedono tecnologie e infrastrutture diverse. Ottimizzare lo stack tecnologico per i propri casi d'uso specifici è fondamentale, ma anche molto complicato. Le soluzioni di cloud pubblico non sempre offrono i livelli di personalizzazione e ottimizzazione delle prestazioni necessari per determinati carichi di lavoro AI.

5. Inferenza e ottimizzazione in prossimità delle origini dati: per ridurre la latenza e migliorare le prestazioni, è fondamentale avvicinare l'analisi ai dati. Le soluzioni di cloud pubblico comportano spesso un trasferimento di dati a grande distanza, che produce ritardi e inefficienze. Eseguendo le operazioni di inferenza e ottimizzazione in posizioni vicine alle origini dati, è possibile aumentare notevolmente la velocità e la precisione dei modelli AI.

Queste problematiche sottolineano l'importanza di valutare le soluzioni di cloud ibrido e privato per l'implementazione dell'AI perché, rispetto alle alternative basate sul cloud pubblico, offrono livelli superiori di controllo, personalizzazione e gestibilità dei costi.

Sintesi dei risultati

HPE Private Cloud AI offre numerose caratteristiche e funzioni chiave per la risoluzione delle principali problematiche affrontate dalle aziende durante la distribuzione dei carichi di lavoro AI:

1. Accelerazione del time to value: HPE accelera lo sviluppo e la distribuzione dell'AI con tool e infrastrutture preintegrati e collaudati, riducendo il time to market.

2. Ripetibilità: la soluzione consente di realizzare distribuzioni AI scalabili, coerenti e ad altre prestazioni, con stack infrastrutturali standardizzati che garantiscono affidabilità ed efficienza.

3. Esperienza cloud: la soluzione HPE promuove la crescita e aumenta la flessibilità grazie alla piattaforma HPE GreenLake e all'infrastruttura scalabile, che consentono di soddisfare esigenze aziendali in continua evoluzione.

4. Sicurezza: le misure di sicurezza complete assicurano la protezione e la compliance dei dati, eliminando le nuove vulnerabilità introdotte con l'adozione dell'AI.

5. Governance e gestione dell'AI: la soluzione semplifica la governance dell'AI unificando access point e controllo, per aumentare la gestibilità negli ambienti ibridi.

6. Partnership strategiche: HPE sfrutta la collaborazione con NVIDIA e gli integratori di sistema per migliorare le capacità e supportare le distribuzioni AI in modo affidabile.

Suggerimenti - Creazione di un business case che privilegia il cloud ibrido/privato al cloud pubblico per l'AI

Le aziende che stanno considerando di implementare l'AI possono ottenere notevoli vantaggi da una soluzione di cloud ibrido/privato. Per identificare accuratamente i casi d'uso dell'AI che generano il massimo valore, occorre comprendere a fondo le esigenze aziendali e avere la capacità di tradurle in progetti AI. Per sviluppare e rendere operativi tali casi d'uso, è necessario integrare i modelli AI nei flussi di lavoro esistenti, in modo da garantire la produzione di risultati aziendali concreti.

È essenziale sfruttare efficacemente i dati nei modelli AI, per generare informazioni approfondite realmente utili senza compromettere la proprietà intellettuale sensibile. Mantenendo un controllo rigoroso sui dati, le aziende possono proteggere gli algoritmi proprietari e le informazioni sensibili da violazioni e accessi non autorizzati, risolvendo i problemi di sicurezza normalmente associati alle soluzioni di cloud pubblico.

È essenziale scalare le iniziative AI per far fronte a esigenze in continuo aumento gestendo efficacemente i costi. Le aziende devono bilanciare le esigenze di potenza di elaborazione e storage con le implicazioni finanziarie, garantendo distribuzioni AI economicamente vantaggiose anche in caso di espansione.

Per garantire il successo, è fondamentale ottimizzare lo stack tecnologico per applicazioni AI specifiche. Le imprese devono personalizzare e ottimizzare le tecnologie in uso per soddisfare le esigenze dei vari carichi di lavoro AI, fornendo le soluzioni ideali per i diversi casi d'uso.

È inoltre importante eseguire le operazioni di inferenza e ottimizzazione dei modelli AI in una posizione vicina all'origine dati. Questo approccio riduce la latenza e aumenta le prestazioni, avvicinando inferenza e ottimizzazione alle origini dati per aumentare la velocità e l'accuratezza dei modelli AI.

Le soluzioni AI basate sul cloud privato offrono numerosi vantaggi chiave alle aziende che hanno l'esigenza di distribuire le soluzioni AI in tutta sicurezza on-premise, nelle strutture in colocation, nei data center o nelle sedi all'edge. La prossimità dei dati accelera l'accesso e l'elaborazione, migliorando l'efficienza e le prestazioni dei modelli AI. La governance completa semplifica la gestione, garantisce affidabilità e aumenta la sicurezza dei carichi di lavoro collaborativi. Il controllo dei costi viene migliorato tramite modelli finanziari flessibili, evitando l'aumento delle spese associato alle soluzioni di cloud pubblico. Inoltre, l'ottimizzazione dello stack tecnologico per casi d'uso specifici assicura la realizzazione di distribuzioni AI su misura, per aumentare al massimo i livelli di impatto ed efficienza.

Conclusioni

L'evoluzione rapida dell'AI generativa presenta sia opportunità che problematiche. Mentre tentano di sfruttare tutto il potenziale dell'AI, le aziende devono destreggiarsi all'interno di ambienti tecnologici complessi, garantire alti livelli di sicurezza e gestire efficacemente sia la scalabilità che i costi. NVIDIA AI Computing by HPE fornisce una soluzione completa a tutte queste problematiche, offrendo un'infrastruttura AI scalabile, sicura e con prestazioni ottimizzate. Sfruttando il cloud privato chiavi in mano per l'AI e le partnership strategiche di HPE, le aziende possono accelerare il loro percorso di adozione dell'IA, velocizzare il time to market e mantenere il controllo su dati e infrastruttura. Con uno sguardo al futuro, HPE intende continuare a impegnarsi per promuovere l'innovazione e aiutare le aziende a liberare il potenziale rivoluzionario dell'AI. Attraverso lo sviluppo continuo e le collaborazioni strategiche, HPE punta a fornire soluzioni all'avanguardia che consentano alle aziende di entrare a testa alta in un futuro basato sull'AI.

Informazioni importanti su questo report

COLLABORATORI

Steven Dickens
Research Analyst | The Futurum Group

EDITORE

Daniel Newman
CEO | The Futurum Group

PER INFORMAZIONI

È possibile contattarci per discutere il presente report. The Futurum Group risponderà tempestivamente.

CITAZIONI

Il presente documento può essere citato da giornalisti e analisti accreditati, ma le citazioni devono essere effettuate in modo contestuale, riportando il nome dell'autore, la relativa qualifica e la dicitura "The Futurum Group". Per le citazioni da parte di altri operatori o analisti non accreditati è necessario il previo consenso scritto di The Futurum Group.

LICENZE

INFORMATIVA

The Futurum Group fornisce ricerche, analisi, opinioni e consulenze a numerose società high-tech, incluse quelle citate nel presente documento. I nostri dipendenti non detengono partecipazioni in alcuna delle società citate nel presente documento.

Hewlett Packard Enterprise

INFORMAZIONI SU HPE

HPE Financial Services fornisce una combinazione di informazioni sulla tecnologia, competenze finanziarie e una spiccata attenzione alla sostenibilità, allo scopo di creare cicli di vita IT più intelligenti per clienti e partner di ogni dimensione. Lavorando sull'intero parco tecnologico, dall'edge al cloud fino all'utente finale, il nostro approccio collaborativo consente di realizzare soluzioni di asset management che, oltre a liberare il capitale e ad aumentare al massimo la capacità, promuovono anche prassi sostenibili coerenti a livello globale. Per ulteriori informazioni, visita il sito: hpe.com/hpefinancialservices

NVIDIA

INFORMAZIONI SU NVIDIA

NVIDIA progetta i chip, i sistemi e i software più avanzati per le fabbriche AI del futuro. Realizziamo nuovi servizi di AI che aiutano le aziende a creare autonomamente le proprie fabbriche AI. Da oltre 30 anni, scienziati, ricercatori, sviluppatori e creativi usano le tecnologie NVIDIA per fare cose straordinarie. Oggi più di 4 milioni di sviluppatori creano migliaia di applicazioni per l'elaborazione accelerata. Le tecnologie AI NVIDIA vengono utilizzate da oltre 40.000 aziende, tra cui 15.000 startup globali nell'ambito di NVIDIA Inception.

The Futurum Group

INFORMAZIONI SU THE FUTURUM GROUP

The Futurum Group è una società di consulenza, ricerca e analisi indipendente che si occupa di innovazione, tecnologie e tendenze rivoluzionarie. Ogni giorno i nostri analisti, ricercatori e consulenti aiutano i leader aziendali di tutto il mondo ad anticipare svolte radicali nei rispettivi settori e a sfruttare l'innovazione dirompente per acquisire o mantenere un vantaggio competitivo sul mercato.

The Futurum Group

CONTATTI

The Futurum Group LLC I futurumgroup.com I (833) 722-5337 I