SPOTLIGHT

Sponsorizzato da: HPE

Le implementazioni dell'IA e della GenAI richiedono prestazioni, scalabilità, semplicità e produttività. Il corretto equilibrio tra risorse e implementazioni consente di migliorare i risultati dell'IA e di ridurre il time to value.

I carichi di lavoro di IA richiedono uno storage dei file dalle prestazioni elevate

Aprile 2024

Autore: Dave Pearson, vicepresidente delle infrastrutture a livello globale

Introduzione

Tradizionalmente, le aziende hanno dovuto gestire priorità contrastanti in termini di investimenti nell'infrastruttura dati, vale a dire massime prestazioni con costi minimi, crescita estrema dei dati e della scalabilità con la massima efficienza possibile, riduzione del time to value in presenza di un aumento del volume dei dati e dipendenze e metodologie di implementazione più complesse rispetto al passato. L'avvento dell'IA predittiva
e interpretativa ha spinto in avanti i confini di questo scenario, mentre la recente e improvvisa esplosione delle iniziative legate all'IA generativa (GenAI) rende ancora più importante l'equilibrio fra queste priorità per gli acquirenti e amministratori IT.

Nel sondaggio *2023 Future of Digital Infrastructure Worldwide Sentiment Survey* (Sondaggio emotivo sul futuro delle infrastrutture digitali a livello globale) di IDC, il secondo KPI in termini di importanza indicato dai dirigenti senior e dai membri dei consigli di amministrazione delle aziende è la tempestività delle informazioni relative ai dati mission-critical (42%), aspetto che dimostra l'importanza critica della trasformazione dell'IA e della GenAI. Tuttavia, i miglioramenti dell'efficienza operativa IT e cloud si trovano ancora in cima all'elenco delle priorità (52%), aspetto che indica la necessità di eliminare le zone isolate delle aziende a livello tecnologico e operativo, migliorare l'accesso ai dati nelle imprese, per utilizzare in modo efficiente le costose risorse di calcolo, e aumentare la produttività dei team di trasformazione dedicati all'IA e alla GenAI.

I dati non strutturati necessari per le iniziative di IA e GenAI impongono criticità specifiche alle infrastrutture dati tradizionali, che offrono prestazioni incentrate soprattutto sull'accesso a blocchi e sull'implementazione di soluzioni specifiche dei diversi carichi di lavoro. La necessità di un'infrastruttura più efficace per lo storage dei dati e file non strutturati appare ormai evidente.

SINTESI

PRINCIPALI CONCLUSIONI

- Uno storage dei file scalabile e dalle prestazioni elevate è fondamentale per la gestione dei 244 ZB di dati non strutturati che si prevede che le aziende creino annualmente entro il 2027.
- I carichi di lavoro dedicati all'IA e alla GenAI dovranno utilizzare implementazioni ibride e multicloud per sfruttare tutti i dati aziendali: se unite a strumenti di gestione unificati tra cloud pubblici e privati, le soluzioni as-a-service e on-premises offrono un'esperienza operativa simile a quella del cloud.
- I carichi di lavoro dedicati all'IA e alla GenAI presentano esigenze fuori dalla portata dei sistemi di storage tradizionali. Le prestazioni, la scalabilità, la semplicità e l'efficienza di tali sistemi possono incrementare la produttività dei team e delle iniziative di trasformazione dell'IA.

Nel 2023, secondo la ricerca Global DataSphere di IDC, il 77% dei dati archiviati sarà di tipo non strutturato (es. immagini, video, testo e audio) e conterrà svariati elementi, come ad esempio file di produttività per l'ufficio, dati medici e sorgenti software (vedi *Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023-2027,* Previsione sui dati strutturati e non strutturati Global DataSphere e Global StorageSphere, IDC n. US50397723 del giugno 2023). La crescita dei dati non strutturati continuerà senza sosta: l'84% dei 291 ZB generati nel 2027 sarà costituito da informazioni di questo tipo. Anche se il testo è ancora l'elemento maggiormente presente nei nuovi dati generati dalla GenAI, secondo le previsioni di IDC, il 75% dei nuovi dati generati entro il 2028 sarà costituito da testo, immagini e video, suddivisi in parti grossomodo uguali.

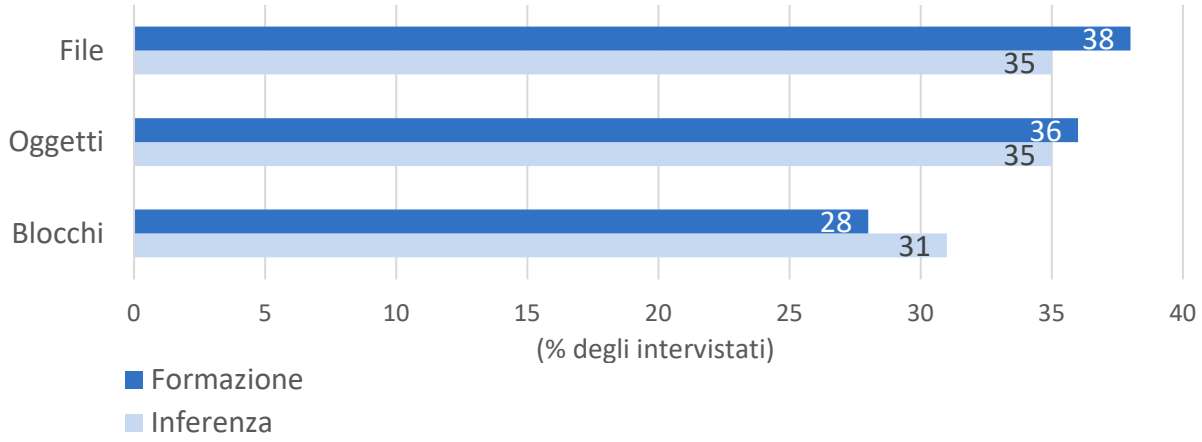
Questi fattori stanno spingendo le organizzazioni a occuparsi delle nuove metodologie di implementazione dell'infrastruttura dati.

- Occorre uno storage dei file efficiente e dalle prestazioni elevate per gestire risorse di calcolo costose e accelerate, aumentare al massimo la produttività e ridurre al minimo il time to value delle iniziative di IA e GenAI.
- Le aziende devono gestire una richiesta variabile e crescente di funzionalità usando implementazioni multicloud, ibride, as-a-service e on-premises per i dati che devono restare nei confini aziendali per motivi di sicurezza, protezione dei dati, privacy, conformità normativa o governance. Circa l'84% dei soggetti intervistati per la nostra indagine 2023 *IT Infrastructure for Storage and Data Management* (Infrastruttura IT per lo storage e la gestione dei dati) ha scelto un approccio basato sul cloud o multicloud ibrido per le implementazioni dell'infrastruttura dati.
- Questo scenario richiede conoscenza delle applicazioni, osservabilità e gestione dell'intera infrastruttura dati. Circa il 70% dei soggetti intervistati per la nostra indagine 2023 *IT Infrastructure for Storage and Data Management* (Infrastruttura IT per lo storage e la gestione dei dati) utilizza, o desidera utilizzare, un sistema di gestione multicloud unificato per l'amministrazione omogenea degli ambienti di storage eterogenei.

Vantaggi

Secondo gli intervistati per l'indagine *AI View* 2023 (Quadro sull'IA) del settembre 2023, lo storage di file, oggetti e blocchi è un elemento fondamentale per l'infrastruttura di IA in termini di accesso ai dati. Lo storage dei file registra una prevalenza per le attività di formazione, così come lo storage a oggetti si dimostra dominante per quanto riguarda l'inferenza (vedi figura 1).

FIGURA 1. ***L'importanza dello storage di file, oggetti e blocchi per l'infrastruttura di AI***



n = 400

Fonte: sondaggio "AI View 2023" di IDC, settembre 2023

Un sistema di storage dei file dalle prestazioni elevate e dalla scalabilità in grado di supportare le richieste variabili e crescenti dei carichi di lavoro dell'IA e della GenAI può favorire l'utilizzo ottimale delle costose risorse di calcolo in tutte le fasi del processo legato a queste due tecnologie. Ciò consente di eliminare i colli di bottiglia dello storage a livello di prestazioni. Tutte le offerte flash NVMe sono dotate delle caratteristiche in termini di throughput e latenza necessarie per numerose fasi dei flussi di lavoro di IA e GenAI. È possibile inserire tutto questo in un singolo livello di storage, in modo da consolidare e ottimizzare la pipeline dei dati per queste attività altamente impegnative. Così facendo, è possibile aumentare la produttività delle risorse più costose, come ad esempio i data scientist e data engineer, e ottimizzare il time to insights e la trasformazione dell'IA.

L'unificazione della gestione, del monitoraggio e della protezione dello storage dei file consente a sviluppatori, data engineer e data scientist di estrarre valore dalle informazioni presenti nei file non strutturati di un namespace globale. Si tratta di dati provenienti da una vasta gamma di origini e modalità di distribuzione, tra cui internamente all'azienda, edge e cloud. Tale risultato è possibile attraverso una console di gestione semplificata che estenda a tutta l'azienda le funzionalità e i servizi legati ai file. L'aumento della produttività e la riduzione del time to value consentono di aumentare al massimo il ROI delle attività di trasformazione dell'IA e della GenAI, favorendo la collaborazione tra personale, team e unità aziendali.

I limiti delle architetture di precedente generazione e le lacune tecniche sono alcuni problemi legati alla complessità che ostacolano la trasformazione digitale. Le aziende devono assolutamente superare queste criticità. Attraverso la semplificazione dell'amministrazione e il superamento della carenza di competenze e risorse nel campo della trasformazione dell'IA, le aziende potranno aumentare la produttività amministrativa e gestire il ciclo di vita, la protezione, la sicurezza, conformità e governance dei dati mediante i servizi dati e risorse non vincolate dai limiti di tempo.

È possibile ottenere risultati come ad esempio aumento del ROI, riduzione del TCO e maggiore sostenibilità grazie all'uso efficiente delle risorse di maggiore densità e capacità e alla presenza di funzionalità adeguate di scalabilità e riduzione dei

dati. Ciò consente di migliorare l'efficienza operativa, le metriche dell'energia e del raffreddamento e i risultati ESG per le parti coinvolte di un'organizzazione.

Prendere in considerazione HPE GreenLake per lo storage dei file

Hewlett Packard Enterprise (HPE) offre servizi infrastrutturali per data center da quasi un decennio, sia agendo in modo indipendente che come divisione di Hewlett-Packard Company (dalla fondazione al 2015). Grazie a un fatturato per l'esercizio 2023 pari a 29,1 miliardi di USD, aumentato del 2% rispetto al 2022, HPE viene considerata, secondo l'Enterprise Storage Systems Tracker di IDC, una delle più importanti realtà del settore dei sistemi di storage esterni, delle soluzioni interne e di quelle basate su server. HPE Alletra Storage, la principale linea di prodotti nel campo dei sistemi di storage di HPE per il 2023, vanta prodotti prestigiosi come ad esempio HPE SimpliVity, HPE Primera e HPE Nimble Storage.

Nel novembre 2017, HPE ha lanciato la piattaforma edge-to-cloud HPE GreenLake, che offre soluzioni infrastrutturali on-premise simili al modello on-demand, o as-a-service, del cloud pubblico. Sin dalle origini, HPE ha ampliato la linea HPE GreenLake fino a raggiungere oltre 50 servizi, che offrono funzioni come ad esempio la gestione ottimale degli ambienti multicloud ibridi estesi su infrastrutture on-premise, edge e cloud pubblici e privati.

HPE GreenLake for File Storage è stato commercializzato nell'aprile 2023 su HPE Alletra Storage MP, una nuova piattaforma hardware, insieme al software VAST Data. Questa nuova combinazione di hardware modulare, interamente flash e dotato della tecnologia NVMe e dell'architettura VAST DASE (Disaggregated Shared Everything, elementi condivisi e completamente disaggregati) per lo storage dei file, offre uno storage dalle alte prestazioni e sull'ordine degli exabyte che consente di evitare i tipici compromessi imposti dalle soluzioni che costringono a dimensionare in modo sinergico le risorse e capacità di calcolo e storage.

Nel marzo 2024, l'offerta è stata rinnovata con una versione dalle prestazioni più elevate che offre netti aumenti in termini di performance, densità e capacità di calcolo e storage. Oltre alle configurazioni standard, l'azienda offre opzioni con maggiori densità di calcolo e storage che offrono una maggiore flessibilità nelle diverse fasi del percorso dei clienti verso l'IA. Alcuni clienti potrebbero essere giunti al limite delle funzionalità delle proprie soluzioni e richiedere quindi maggiori prestazioni per attività come carichi di lavoro dell'IA e GenAI, machine learning (ML), calcolo ad alte prestazioni (HPC), big data, scienze biologiche, analisi finanziaria, media e intrattenimento, data lake e implementazioni degli archivi ad accesso rapido.

HPE GreenLake for File Storage è una soluzione di storage dalle prestazioni elevate, adatta a diversi carichi di lavoro e che consente di scegliere fra i modelli tradizionali per l'acquisto e la licenza dello storage o un approccio a consumo basato sulle funzionalità di GreenLake.

Le caratteristiche di gestione, monitoraggio e protezione centralizzate della piattaforma HPE GreenLake semplificano il lavoro di amministrazione degli ambienti ibridi multicloud. La configurazione semplificata con namespace globale agevola la collaborazione e le implementazioni, il rilevamento automatico, l'onboarding e la messa a punto dei dispositivi. HPE GreenLake offre una suite di servizi per i dati in grado di migliorare la produttività dei data engineer e data scientist attraverso l'eliminazione di gran parte dei costi di gestione associati alle architetture di precedente generazione. Si tratta soprattutto di aspetti come ad esempio gestione del ciclo di vita, backup e ripristino d'emergenza, aggiornamenti

senza interruzioni e manutenzione proattiva, che vengono gestiti attraverso la console unica di HPE GreenLake.

La sostenibilità e l'efficienza operativa sono elementi fondamentali per ottenere un ROI adeguato per le attività di IA e GenAI. La scalabilità indipendente delle prestazioni e della capacità tramite l'aggiunta flessibile e separata dei nodi di controllo o capacità evita il sottoutilizzo delle risorse principali, un problema tipico dei sistemi tradizionali che richiedono diversi aggiornamenti per ottenere una scalabilità progressiva. L'aumento della densità delle risorse di storage attraverso supporti flash ottimizzati in funzione della capacità, insieme alle tecnologie di riduzione dei dati, consente di ridurre l'ingombro, le emissioni inquinanti e i requisiti di alimentazione e raffreddamento dei sistemi di storage negli ambienti che non riescono a supportare al meglio le risorse di calcolo con accelerazione della GPU e dall'elevato consumo di energia. HPE sfrutta l'algoritmo di riduzione dei dati Similarity, disponibile in HPE GreenLake for File Storage, per la riduzione dei dati granulari nel namespace globale. Questo approccio unisce gli aspetti più efficaci della compressione e deduplica senza sovraccaricare i sistemi di storage. HPE GreenLake for File Storage supporta Infiniband, NVIDIA GPUDirect e l'accesso diretto e remoto alla memoria, in modo da ottimizzare l'utilizzo della GPU, ridurre il tempo per la generazione degli approfondimenti e aumentare il ROI per le iniziative critiche di IA e GenAI.

L'intersezione tra la richiesta di maggiori prestazioni e capacità per i carichi di lavoro di IA ad alta intensità di dati e il controllo delle informazioni più preziose apre un'opportunità per lo storage dei file scalabile e dalle prestazioni elevate.

Sfide

Gli utenti alle prime armi con l'IA e la GenAI iniziano spesso il proprio percorso di trasformazione con i servizi cloud di IA, che costituiscono gran parte del mercato attuale. Tuttavia, la progressiva maturazione delle organizzazioni spinge a lasciare i dati di maggiore importanza sotto il controllo delle imprese, ovvero all'interno del perimetro aziendale. Operare in questa modalità di cloud ibrido o multicloud può aumentare la complessità di questi ambienti, a meno che i fornitori non riescano a garantire un carico di lavoro e una mobilità dei dati senza attriti, nonché osservabilità e gestione all'interno dell'organizzazione. L'offerta di tecnologie, servizi e risorse di alto livello in modo diretto o attraverso ecosistemi di partner, in modo da velocizzare il raggiungimento della maturità degli utenti di IA e GenAI, è un requisito fondamentale per il successo di queste iniziative ad alto valore.

HPE non è l'unica azienda a proporre offerte as-a-service nel settore dello storage. Pertanto, dovrà contrastare una forte concorrenza da parte dei produttori che intendono supportare i carichi di lavoro di IA e GenAI attraverso una vasta gamma di tecnologie e implementazioni hardware e software. HPE dovrà espandere la proposta a valore dell'ecosistema GreenLake e l'importanza di elementi come prestazioni, semplicità ed efficienza per l'IA. Numerosi clienti alle prime armi nel campo dell'IA sono in cerca di semplici servizi cloud o soluzioni full-stack "chiavi in mano", poiché non dispongono di risorse qualificate in questo campo. HPE GreenLake for File Storage dovrà accelerare il percorso di questi clienti verso la maturità dell'IA e offrire un maggiore valore nel ciclo di vita di queste iniziative.

Conclusione

L'aumento del volume dei dati non strutturati resta un problema per le aziende alle prese con ogni fase della trasformazione digitale e del lavoro con l'IA. Aggiungendo la necessità di ottenere informazioni tempestive dalle analisi avanzate, dall'IA e dalla GenAI, la crescita delle architetture di storage on-demand e a scalabilità orizzontale continuerà

senza sosta almeno per i prossimi cinque anni. Le aziende, che in passato sceglievano il cloud pubblico grazie alla rapidità di implementazione e scalabilità delle capacità di storage, si vedono costrette ad archiviare parte dei dati nei propri data center, uno scenario che richiede alternative interne volte a semplificare l'acquisizione, l'implementazione, il funzionamento e l'espansione dello storage dei dati non strutturati.

HPE GreenLake for File Storage viene implementato su una piattaforma hardware unificata, ad alte prestazioni e progettata per la gestione dei carichi di lavoro a elevata intensità di dati. Grazie all'aumento delle prestazioni e delle capacità offerte, HPE propone una soluzione di storage dei file dalle performance elevate e adatta a carichi di lavoro molto pesanti che consente di scegliere i modelli di acquisto e licenza dello storage tradizionali o un approccio a consumo mediante HPE GreenLake, in grado di sfruttare i paradigmi operativi e gestionali dei servizi cloud. L'intersezione tra la richiesta di maggiori prestazioni e capacità per i carichi di lavoro di IA ad alta intensità di dati e il controllo delle informazioni più preziose apre un'opportunità per lo storage dei file scalabile e dalle prestazioni elevate. Il continuo interesse verso le iniziative di IA, GenAI, ML e HPC dovrebbe offrire ai fornitori chiare opportunità di crescita nel campo di questi carichi di lavoro emergenti.

Informazioni sull'analista

Dave Pearson, *vicepresidente del reparto di ricerca sulle infrastrutture a livello mondiale*

Dave Pearson è il vicepresidente della ricerca sulle infrastrutture a livello mondiale e il direttore degli studi globali del reparto dedicato allo storage e ai sistemi convergenti di IDC. Dave e il suo team forniscono approfondimenti a livello globale sugli andamenti tecnologici relativi a storage, infrastrutture integrate, iperconvergenti e componibili, nonché sull'adozione da parte del mercato. Questi materiali comprendono lo storage per le applicazioni di calcolo a elevate prestazioni, come ad esempio HPC, IA e analisi.

MESSAGGIO DELLO SPONSOR

HPE GreenLake for File Storage è uno storage aziendale a scalabilità orizzontale per i file che accelera la crescita dei carichi di lavoro per l'IA ad alta intensità di dati. Questo servizio cloud offre prestazioni di livello aziendale, semplicità ed efficienza in funzione dell'IA. HPE GreenLake for File Storage offre un maggiore valore dai dati, tempi più rapidi per l'analisi e il conseguimento dei vantaggi competitivi, permette ai data scientist e ai proprietari delle applicazioni della LOB di aumentare la produttività e concentrarsi sul rilevamento e sull'innovazione e garantisce un'elevata sostenibilità grazie a ottimi risultati in termini di densità della capacità, ingombro dei data center, efficienza energetica e del raffreddamento e riduzione delle emissioni inquinanti.

Scopri in che modo velocizzare il tuo percorso verso l'IA con la soluzione di storage dei file più completa del settore, che consente di archiviare, gestire e proteggere i dati nel cloud ibrido avvalendosi delle prestazioni, della semplicità e dell'efficienza richieste dalle applicazioni di IA.

Guarda il video dimostrativo.

IDC Custom Solutions

Il contenuto del presente documento è adattato da una ricerca IDC disponibile all'indirizzo www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T. +1 508.872.8200
F. +1 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com

Questa pubblicazione è prodotta da IDC Custom Solutions. Le opinioni, le analisi e i risultati della ricerca indicati nel presente documento derivano da uno studio e un'analisi condotti e pubblicati in maniera indipendente da IDC. Le eventuali sponsorizzazioni verranno indicate esplicitamente. IDC Custom Solutions distribuisce i contenuti di IDC in numerosi formati adatti alle esigenze aziendali. La licenza per la distribuzione dei contenuti non implica l'approvazione del licenziatario da parte di IDC, né che IDC possegga una qualsiasi opinione su tale soggetto.

Pubblicazione esterna di informazioni e dati di IDC. L'utilizzo di qualsiasi informazione IDC presente all'interno di pubblicità, comunicati stampa o materiale promozionale è soggetta all'approvazione scritta da parte del vicepresidente o responsabile nazionale di IDC appropriato. Allegare alla richiesta la bozza del documento proposto. IDC si riserva il diritto di negare l'approvazione dell'utilizzo esterno del proprio materiale per qualsiasi motivo.

Copyright 2024 IDC. La riproduzione senza autorizzazione scritta è severamente vietata.