Intelligenza artificiale generativa e valore NetApp

NetApp artificial intelligence solutions

NetApp
December 04, 2025

Sommario

Intelligenza artificiale generativa e valore NetApp

La domanda di intelligenza artificiale (IA) generativa sta determinando cambiamenti radicali in tutti i settori, migliorando la creatività aziendale e l'innovazione dei prodotti.

Astratto

Molte organizzazioni utilizzano l'intelligenza artificiale generativa per sviluppare nuove funzionalità di prodotto, migliorare la produttività ingegneristica e prototipare applicazioni basate sull'intelligenza artificiale che offrono risultati ed esperienze migliori per i consumatori. L'intelligenza artificiale generativa, come i Generative Pre-trained Transformers (GPT), utilizza reti neurali per creare nuovi contenuti, tra cui testo, audio e video. Considerata l'estrema scala e gli enormi set di dati coinvolti nei modelli linguistici di grandi dimensioni (LLM), è fondamentale progettare un'infrastruttura di intelligenza artificiale solida che sfrutti le interessanti funzionalità di archiviazione dei dati delle opzioni di distribuzione on-premise, ibride e multicloud e riduca i rischi associati alla mobilità dei dati, alla protezione dei dati e alla governance prima che le aziende possano progettare soluzioni di intelligenza artificiale. In questo documento vengono descritte queste considerazioni e le corrispondenti funzionalità di intelligenza artificiale NetApp che consentono una gestione e uno spostamento dei dati senza interruzioni lungo la pipeline dei dati di intelligenza artificiale per l'addestramento, il riaddestramento, la messa a punto e l'inferenza dei modelli di intelligenza artificiale generativa.

Sintesi

Più di recente, dopo il lancio di ChatGPT, uno spin-off di GPT-3 nel novembre 2022, i nuovi strumenti di intelligenza artificiale utilizzati per generare testo, codice, immagini o persino proteine terapeutiche in risposta alle richieste degli utenti hanno acquisito notevole fama. Ciò indica che gli utenti possono effettuare una richiesta utilizzando il linguaggio naturale e l'intelligenza artificiale interpreterà e genererà testo, come articoli di giornale o descrizioni di prodotti che riflettono la richiesta dell'utente o produrranno codice, musica, parlato, effetti visivi e risorse 3D utilizzando algoritmi addestrati su dati già esistenti. Di conseguenza, espressioni come Diffusione stabile, Allucinazioni, Ingegneria rapida e Allineamento del valore stanno emergendo rapidamente nella progettazione dei sistemi di intelligenza artificiale. Questi modelli di apprendimento automatico (ML) auto-supervisionati o semi-supervisionati stanno diventando ampiamente disponibili come modelli di base pre-addestrati (FM) tramite fornitori di servizi cloud e altri fornitori di aziende di intelligenza artificiale, e vengono adottati da varie aziende in tutti i settori per un'ampia gamma di attività NLP (elaborazione del linguaggio naturale) a valle. Come affermato da società di analisi di ricerca come McKinsey: "L'impatto dell'intelligenza artificiale generativa sulla produttività potrebbe aggiungere migliaia di miliardi di dollari di valore all'economia globale". Mentre le aziende stanno ripensando l'intelligenza artificiale come partner intellettuale degli esseri umani e i gestori di gestione stanno ampliando contemporaneamente ciò che aziende e istituzioni possono fare con l'intelligenza artificiale generativa, le opportunità di gestire enormi volumi di dati continueranno a crescere. Questo documento presenta informazioni introduttive sull'intelligenza artificiale generativa e sui concetti di progettazione in relazione alle funzionalità NetApp che apportano valore ai clienti NetApp , sia in ambienti locali che ibridi o multicloud.

Quindi, cosa ci guadagnano i clienti utilizzando NetApp nei loro ambienti di intelligenza artificiale?

NetApp aiuta le organizzazioni a far fronte alle complessità create dalla rapida crescita dei dati e del cloud, dalla gestione multi-cloud e dall'adozione di tecnologie di nuova generazione, come l'intelligenza artificiale. NetApp ha combinato diverse funzionalità in un software di gestione dati intelligente e in un'infrastruttura di storage ben bilanciata con prestazioni elevate ottimizzate per i carichi di lavoro di intelligenza artificiale. Le soluzioni di intelligenza artificiale generativa come gli LLM devono leggere ed elaborare i set di dati sorgente dall'archiviazione alla memoria numerose volte per promuovere l'intelligenza.

NetApp è leader nelle tecnologie di mobilità, governance e sicurezza dei dati nell'ecosistema edge-to-core-to-cloud, aiutando i clienti aziendali a creare soluzioni di intelligenza artificiale su larga scala. NetApp, grazie a una solida rete di partner, ha aiutato i responsabili dei dati, gli ingegneri dell'intelligenza artificiale, gli architetti aziendali e gli scienziati dei dati a progettare una pipeline di dati fluida per la preparazione dei dati, la protezione dei dati e le responsabilità di gestione strategica dei dati per l'addestramento e l'inferenza dei modelli di intelligenza artificiale, ottimizzando le prestazioni e la scalabilità del ciclo di vita di intelligenza artificiale/apprendimento automatico. Le tecnologie e le funzionalità dei dati NetApp , come NetApp ONTAP AI per la pipeline di dati di deep learning, NetApp SnapMirror per il trasporto fluido ed efficiente dei dati tra endpoint di storage e NetApp FlexCache per il rendering in tempo reale quando il flusso di dati passa da batch a tempo reale e l'ingegneria dei dati avviene al momento opportuno, apportano valore all'implementazione di modelli di intelligenza artificiale generativa in tempo reale. Con l'adozione di nuovi strumenti di intelligenza artificiale da parte delle aziende di ogni tipo, si trovano ad affrontare sfide legate ai dati, dall'edge al data center fino al cloud, che richiedono soluzioni di intelligenza artificiale scalabili, responsabili e spiegabili.

In qualità di autorità in materia di dati su cloud ibrido e multi-cloud, NetApp si impegna a creare una rete di partner e soluzioni congiunte che possano aiutare in tutti gli aspetti della costruzione di una pipeline di dati e di data lake per l'addestramento di modelli di intelligenza artificiale generativa (pre-addestramento), la messa a punto, l'inferenza basata sul contesto e il monitoraggio del decadimento del modello di LLM.

Cos'è l'intelligenza artificiale generativa?

L'intelligenza artificiale generativa sta cambiando il modo in cui creiamo contenuti, generiamo nuovi concetti di design ed esploriamo composizioni innovative. Illustra framework di reti neurali come Generative Adversarial Network (GAN), Variational Autoencoders (VAE) e Generative Pre-Trained Transformers (GPT), che possono generare nuovi contenuti come testo, codice, immagini, audio, video e dati sintetici. Modelli basati su trasformatori come Chat-GPT di OpenAI, Bard di Google, BLOOM di Hugging Face e LLaMA di Meta sono emersi come tecnologie fondamentali alla base di molti progressi nei modelli linguistici di grandi dimensioni. Allo stesso modo, Dall-E di OpenAI, CM3leon di Meta e Imagen di Google sono esempi di modelli di diffusione testo-immagine che offrono ai clienti un livello di fotorealismo senza precedenti per creare nuove immagini complesse da zero o modificare immagini esistenti per generare immagini contestuali di alta qualità utilizzando l'aumento del set di dati e la sintesi testo-immagine che collega la semantica testuale e visiva. Gli artisti digitali stanno iniziando ad applicare una combinazione di tecnologie di rendering come NeRF (Neural Radiance Field) con l'intelligenza artificiale generativa per convertire immagini 2D statiche in scene 3D immersive. In generale, gli LLM sono ampiamente caratterizzati da quattro parametri: (1) Dimensione del modello (tipicamente in miliardi di parametri); (2) Dimensione del set di dati di addestramento; (3) Costo dell'addestramento e (4) Prestazioni del modello dopo l'addestramento. Anche gli LLM rientrano principalmente in tre architetture di trasformatori. (i) Modelli solo con encoder. Ad esempio BERT (Google, 2018); (ii) modelli Encoder-Decoder, ad esempio BART (Meta, 2020) e (iii) modelli solo Decoder. Ad esempio LLaMA (Meta, 2023), PaLM-E (Google, 2023). A seconda dei requisiti aziendali, indipendentemente dall'architettura scelta dall'azienda, il numero di parametri del modello (N) e il numero di token (D) nel set di dati di addestramento determinano generalmente il costo di base dell'addestramento (pre-addestramento) o della messa a punto di un LLM.

Casi d'uso aziendali e attività NLP a valle

Le aziende di tutti i settori stanno scoprendo sempre più il potenziale dell'intelligenza artificiale nell'estrarre e produrre nuove forme di valore dai dati esistenti per le operazioni aziendali, le vendite, il marketing e i servizi legali. Secondo le informazioni di mercato di IDC (International Data Corporation) sui casi d'uso e gli investimenti globali nell'intelligenza artificiale generativa, la gestione della conoscenza nello sviluppo del software e nella progettazione dei prodotti sarà quella maggiormente interessata, seguita dalla creazione di storyline per il marketing e dalla generazione di codice per gli sviluppatori. Nel settore sanitario, le organizzazioni di ricerca clinica stanno aprendo nuove strade alla medicina. I modelli preaddestrati come ProteinBERT incorporano annotazioni Gene Ontology (GO) per progettare rapidamente strutture proteiche per farmaci, rappresentando una pietra miliare significativa nella scoperta di farmaci, nella bioinformatica e nella

biologia molecolare. Le aziende biotecnologiche hanno avviato sperimentazioni sull'uomo per la medicina generativa scoperta dall'intelligenza artificiale, che mira a curare malattie come la fibrosi polmonare (IPF), una malattia polmonare che provoca cicatrici irreversibili nel tessuto polmonare.

Figura 1: Casi d'uso che guidano l'intelligenza artificiale generativa

[Figura 1: Casi d'uso che guidano l'intelligenza artificiale generativa]

L'aumento dell'adozione dell'automazione, favorito dall'intelligenza artificiale generativa, sta modificando anche l'offerta e la domanda di attività lavorative per molte professioni. Secondo McKinsey, il mercato del lavoro statunitense (diagramma sotto) ha attraversato una rapida transizione, che potrebbe proseguire solo se si considera l'impatto dell'intelligenza artificiale.

Fonte: McKinsey & Company

[Figura 2: Fonte: McKinsey Company]

Il ruolo dello storage nell'intelligenza artificiale generativa

Gli LLM si basano in larga parte su apprendimento profondo, GPU e calcolo. Tuttavia, quando il buffer della GPU si riempie, i dati devono essere scritti rapidamente nella memoria. Mentre alcuni modelli di intelligenza artificiale sono sufficientemente piccoli da poter essere eseguiti in memoria, gli LLM richiedono IOPS elevati e un'archiviazione ad alta velocità per fornire un accesso rapido a grandi set di dati, soprattutto se si tratta di miliardi di token o milioni di immagini. Per un tipico requisito di memoria GPU di un LLM, la memoria necessaria per addestrare un modello con 1 miliardo di parametri potrebbe arrivare fino a 80 GB a 32 bit di precisione completa. In tal caso, LLaMA 2 di Meta, una famiglia di LLM con una scala che va da 7 miliardi a 70 miliardi di parametri, potrebbe richiedere 70x80, ovvero circa 5600 GB o 5,6 TB di RAM GPU. Inoltre, la quantità di memoria necessaria è direttamente proporzionale al numero massimo di token che si desidera generare. Ad esempio, se si desidera generare output fino a 512 token (circa 380 parole), è necessario"512 MB" . Potrebbe sembrare irrilevante, ma se si vogliono produrre lotti più grandi, la cosa inizia a farsi sentire. Ciò rende molto costoso per le organizzazioni addestrare o perfezionare gli LLM in memoria, rendendo così l'archiviazione un elemento fondamentale per l'intelligenza artificiale generativa.

Tre approcci principali agli LLM

Per la maggior parte delle aziende, in base alle tendenze attuali, l'approccio all'implementazione degli LLM può essere condensato in 3 scenari di base. Come descritto in un recente"Harvard Business Review" articolo: (1) Formazione (pre-formazione) di un LLM da zero: costosa e richiede competenze specialistiche in AI/ML; (2) Perfezionamento di un modello di base con dati aziendali: complesso, ma fattibile; (3) Utilizzo della generazione aumentata dal recupero (RAG) per interrogare repository di documenti, API e database vettoriali che contengono dati aziendali. Ciascuno di questi presenta dei compromessi tra impegno, velocità di iterazione, efficienza dei costi e accuratezza del modello nelle loro implementazioni, utilizzati per risolvere diversi tipi di problemi (diagramma seguente).

Figura 3: Tipi di problemi

[Figura 3: Tipi di problemi]

Modelli di fondazione

Un modello di base (FM), noto anche come modello di base, è un modello di intelligenza artificiale (LLM) di grandi dimensioni addestrato su grandi quantità di dati non etichettati, utilizzando l'autosupervisione su larga scala, generalmente adattato per un'ampia gamma di attività NLP a valle. Poiché i dati di addestramento non sono etichettati dagli esseri umani, il modello emerge anziché essere codificato in modo esplicito. Ciò significa

che il modello può generare storie o una narrazione propria senza essere programmato esplicitamente per farlo. Pertanto una caratteristica importante della FM è l'omogeneizzazione, ovvero lo stesso metodo viene utilizzato in molti domini. Tuttavia, grazie alle tecniche di personalizzazione e di messa a punto, i FM integrati nei prodotti che compaiono oggigiorno non solo sono efficaci nella generazione di testo, nella conversione di testo in immagini e nella conversione di testo in codice, ma anche nella spiegazione di attività specifiche di un dominio o nel debug del codice. Ad esempio, FM come Codex di OpenAI o Code Llama di Meta possono generare codice in più linguaggi di programmazione basandosi su descrizioni in linguaggio naturale di un'attività di programmazione. Questi modelli sono competenti in oltre una dozzina di linguaggi di programmazione, tra cui Python, C#, JavaScript, Perl, Ruby e SQL. Comprendono l'intento dell'utente e generano codice specifico che realizza l'attività desiderata, utile per lo sviluppo del software, l'ottimizzazione del codice e l'automazione delle attività di programmazione.

Ottimizzazione, specificità del dominio e riaddestramento

Una delle pratiche più comuni nell'implementazione di LLM dopo la preparazione e la pre-elaborazione dei dati è quella di selezionare un modello pre-addestrato, ovvero addestrato su un set di dati ampio e diversificato. Nel contesto della messa a punto, questo può essere un modello di linguaggio di grandi dimensioni open source come"Meta's Llama 2" addestrato su 70 miliardi di parametri e 2 trilioni di token. Una volta selezionato il modello pre-addestrato, il passo successivo è perfezionarlo sui dati specifici del dominio. Ciò comporta la regolazione dei parametri del modello e l'addestramento sui nuovi dati per adattarlo a un dominio e a un'attività specifici. Ad esempio, BloombergGPT, un LLM proprietario formato su un'ampia gamma di dati finanziari al servizio del settore finanziario.

I modelli specifici per dominio, progettati e addestrati per un compito specifico, presentano generalmente maggiore accuratezza e prestazioni nel loro ambito, ma una bassa trasferibilità ad altri compiti o domini. Quando l'ambiente aziendale e i dati cambiano nel corso di un periodo, l'accuratezza delle previsioni del FM potrebbe iniziare a diminuire rispetto alle prestazioni durante i test. È in questi casi che diventa fondamentale riqualificare o perfezionare il modello.

Il riaddestramento del modello nell'intelligenza artificiale/apprendimento automatico tradizionale si riferisce all'aggiornamento di un modello di apprendimento automatico distribuito con nuovi dati, generalmente eseguito per eliminare due tipi di derive che si verificano. (1) Deriva concettuale: quando il collegamento tra le variabili di input e le variabili di destinazione cambia nel tempo, poiché la descrizione di ciò che vogliamo prevedere cambia, il modello può produrre previsioni imprecise. (2) Deriva dei dati: si verifica quando le caratteristiche dei dati di input cambiano, come cambiamenti nelle abitudini o nel comportamento dei clienti nel tempo e quindi l'incapacità del modello di rispondere a tali cambiamenti.

Allo stesso modo, la riqualificazione si applica ai FM/LLM, ma può essere molto più costosa (in milioni di dollari), quindi non è qualcosa che la maggior parte delle organizzazioni potrebbe prendere in considerazione. È oggetto di ricerca attiva e sta ancora emergendo nel campo degli LLMOps. Quindi, invece di riaddestrare, quando si verifica un decadimento del modello in FM ottimizzati, le aziende potrebbero optare per un'ulteriore ottimizzazione (molto più economica) con un set di dati più recente. Per una prospettiva sui costi, di seguito è riportato un esempio di tabella modello-prezzo di Azure-OpenAI Services. Per ogni categoria di attività, i clienti possono perfezionare e valutare i modelli su set di dati specifici.

Fonte: Microsoft Azure

[Fonte: Microsoft Azure]

Ingegneria e inferenza rapide

Con "prompt engineering" si intendono i metodi efficaci per comunicare con gli LLM per eseguire le attività desiderate senza aggiornare i pesi del modello. Per quanto l'addestramento e la messa a punto dei modelli di intelligenza artificiale siano importanti per le applicazioni NLP, altrettanto importante è l'inferenza, in cui i

modelli addestrati rispondono alle richieste dell'utente. I requisiti di sistema per l'inferenza sono generalmente molto più orientati alle prestazioni di lettura del sistema di archiviazione AI che invia i dati dagli LLM alle GPU, poiché deve essere in grado di applicare miliardi di parametri del modello archiviati per produrre la risposta migliore.

LLMOps, monitoraggio dei modelli e Vectorstores

Analogamente alle tradizionali operazioni di Machine Learning Ops (MLOps), anche le operazioni su modelli di linguaggio di grandi dimensioni (LLMOps) richiedono la collaborazione di data scientist e ingegneri DevOps con strumenti e best practice per la gestione degli LLM negli ambienti di produzione. Tuttavia, il flusso di lavoro e lo stack tecnologico per gli LLM potrebbero variare in alcuni aspetti. Ad esempio, le pipeline LLM create utilizzando framework come LangChain uniscono più chiamate API LLM a endpoint di incorporamento esterni come vectorstore o database vettoriali. L'utilizzo di un endpoint di incorporamento e di un vectorstore per i connettori downstream (come per un database vettoriale) rappresenta uno sviluppo significativo nel modo in cui i dati vengono archiviati e accessibili. A differenza dei tradizionali modelli ML sviluppati da zero, gli LLM spesso si basano sull'apprendimento per trasferimento, poiché questi modelli partono da modelli di apprendimento automatico (FM) che vengono perfezionati con nuovi dati per migliorare le prestazioni in un dominio più specifico. Pertanto, è fondamentale che gli LLMOps forniscano le capacità di gestione del rischio e di monitoraggio del decadimento del modello.

Rischi ed etica nell'era dell'intelligenza artificiale generativa

"ChatGPT: è ingegnoso ma continua a vomitare assurdità." – MIT Tech Review. Il problema dell'intrusione e dell'eliminazione dei rifiuti è sempre stato il problema più spinoso dell'informatica. L'unica differenza con l'intelligenza artificiale generativa è che quest'ultima è in grado di rendere i dati estremamente credibili, il che porta a risultati imprecisi. Gli LLM tendono a inventare fatti per adattarli alla narrazione che stanno costruendo. Pertanto, le aziende che vedono nell'intelligenza artificiale generativa una grande opportunità per ridurre i costi con equivalenti di intelligenza artificiale devono individuare in modo efficiente i deep fake, ridurre i pregiudizi e i rischi per mantenere i sistemi onesti ed etici. Un flusso di dati libero con una solida infrastruttura di intelligenza artificiale che supporti la mobilità dei dati, la qualità dei dati, la governance dei dati e la protezione dei dati tramite crittografia end-to-end e protezioni di intelligenza artificiale è fondamentale nella progettazione di modelli di intelligenza artificiale generativa responsabili e spiegabili.

Scenario del cliente e NetApp

Figura 3: Flusso di lavoro del modello di apprendimento automatico/linguaggio di grandi dimensioni

[Figura 3: Flusso di lavoro del modello di apprendimento automatico/linguaggio di grandi dimensioni]

Stiamo facendo formazione o perfezionando? La questione se (a) addestrare un modello LLM da zero, perfezionare un FM pre-addestrato o utilizzare RAG per recuperare dati da repository di documenti al di fuori di un modello di base e ampliare i prompt e (b) sfruttare LLM open source (ad esempio, Llama 2) o FM proprietari (ad esempio, ChatGPT, Bard, AWS Bedrock) è una decisione strategica per le organizzazioni. Ogni approccio presenta un compromesso tra efficienza dei costi, gravità dei dati, operazioni, accuratezza del modello e gestione degli LLM.

NetApp , in quanto azienda, abbraccia l'intelligenza artificiale internamente, nella sua cultura aziendale e nel suo approccio alla progettazione e all'ingegneria dei prodotti. Ad esempio, la protezione autonoma contro i ransomware di NetApp è realizzata utilizzando l'intelligenza artificiale e l'apprendimento automatico. Fornisce il rilevamento precoce delle anomalie del file system per aiutare a identificare le minacce prima che incidano sulle operazioni. In secondo luogo, NetApp utilizza l'intelligenza artificiale predittiva per le sue operazioni aziendali, come le previsioni di vendita e inventario, e chatbot per assistere i clienti nei servizi di supporto ai prodotti del call center, nelle specifiche tecniche, nella garanzia, nei manuali di assistenza e altro ancora. In terzo luogo, NetApp apporta valore al cliente nella pipeline di dati AI e nel flusso di lavoro ML/LLM tramite

prodotti e soluzioni al servizio dei clienti che creano soluzioni di AI predittiva come la previsione della domanda, l'imaging medico, l'analisi del sentiment e soluzioni di AI generativa come GAN per il rilevamento di anomalie nelle immagini industriali nel settore manifatturiero e l'antiriciclaggio e il rilevamento delle frodi nei servizi bancari e finanziari con prodotti e funzionalità NetApp come NetApp ONTAP AI, NetApp SnapMirror e NetApp FlexCache.

Capacità NetApp

Lo spostamento e la gestione dei dati nelle applicazioni di intelligenza artificiale generativa, come chatbot, generazione di codice, generazione di immagini o espressione di modelli genomici, possono estendersi all'edge, al data center privato e all'ecosistema multicloud ibrido. Ad esempio, un bot di intelligenza artificiale in tempo reale che aiuta un passeggero ad aggiornare il suo biglietto aereo alla classe business tramite un'app per utenti finali esposta tramite API di modelli pre-addestrati come ChatGPT non può svolgere tale compito da solo, poiché le informazioni sul passeggero non sono disponibili pubblicamente su Internet. L'API richiede l'accesso alle informazioni personali del passeggero e alle informazioni sul biglietto della compagnia aerea, che può esistere in un ecosistema ibrido o multicloud. Uno scenario simile potrebbe applicarsi agli scienziati che condividono una molecola di farmaco e i dati dei pazienti tramite un'applicazione per l'utente finale che utilizza LLM per realizzare sperimentazioni cliniche nell'ambito della scoperta di farmaci che coinvolgono istituti di ricerca biomedica one-to-many. I dati sensibili trasmessi ai FM o agli LLM possono includere informazioni personali identificabili (PII), informazioni finanziarie, informazioni sanitarie, dati biometrici, dati sulla posizione, dati sulle comunicazioni, comportamento online e informazioni legali. In un evento di rendering in tempo reale, esecuzione rapida e inferenza edge, si verifica uno spostamento di dati dall'app dell'utente finale agli endpoint di archiviazione tramite modelli LLM open source o proprietari a un data center in sede o a piattaforme cloud pubbliche. In tutti questi scenari, la mobilità e la protezione dei dati sono fondamentali per le operazioni di intelligenza artificiale che coinvolgono LLM, che si basano su grandi set di dati di formazione e sullo spostamento di tali dati.

Figura 4: IA generativa - Pipeline dati LLM

[Figura 4: Pipeline di dati AI-LLM generativi]

Il portafoglio di infrastrutture di storage, dati e servizi cloud di NetApp è basato su un software di gestione dati intelligente.

Preparazione dei dati: il primo pilastro dello stack tecnologico LLM è in gran parte invariato rispetto al vecchio stack ML tradizionale. La pre-elaborazione dei dati nella pipeline dell'IA è necessaria per normalizzare e ripulire i dati prima dell'addestramento o della messa a punto. Questa fase include connettori per l'acquisizione di dati ovunque si trovino, sotto forma di livello Amazon S3 o in sistemi di archiviazione locali, come un archivio file o un archivio oggetti come NetApp StorageGRID.

- NetApp ONTAP* è la tecnologia fondamentale su cui si fondano le soluzioni di storage critiche di NetApp nei data center e nel cloud. ONTAP include diverse funzionalità e capacità di gestione e protezione dei dati, tra cui la protezione automatica contro i ransomware e gli attacchi informatici, funzionalità integrate di trasporto dati e capacità di efficienza di archiviazione per una vasta gamma di architetture, da quelle on-premise, ibride, multicloud in NAS, SAN, oggetti e situazioni di archiviazione definita dal software (SDS) delle distribuzioni LLM.
- NetApp ONTAP AI* per la formazione di modelli di deep learning. NetApp ONTAP supporta NVIDIA GPU Direct Storage con l'uso di NFS su RDMA per i clienti NetApp con cluster di storage ONTAP e nodi di elaborazione NVIDIA DGX. Offre prestazioni convenienti per leggere ed elaborare i set di dati sorgente dall'archiviazione alla memoria più volte, favorendo l'intelligence e consentendo alle organizzazioni di accedere ai LLM con formazione, messa a punto e scalabilità.
- NetApp FlexCache* è una funzionalità di caching remoto che semplifica la distribuzione dei file e memorizza nella cache solo i dati letti attivamente. Ciò può essere utile per la formazione, la

riqualificazione e la messa a punto LLM, offrendo valore ai clienti con requisiti aziendali quali rendering in tempo reale e inferenza LLM.

- NetApp SnapMirror* è una funzionalità ONTAP che replica gli snapshot dei volumi tra due sistemi ONTAP qualsiasi. Questa funzionalità trasferisce in modo ottimale i dati dall'edge al data center locale o al cloud. SnapMirror può essere utilizzato per spostare dati in modo sicuro ed efficiente tra cloud locali e hyperscaler, quando i clienti desiderano sviluppare intelligenza artificiale generativa in cloud con RAG contenenti dati aziendali. Trasferisce in modo efficiente solo le modifiche, risparmiando larghezza di banda e velocizzando la replicazione, apportando così funzionalità essenziali di mobilità dei dati durante le operazioni di formazione, riqualificazione e messa a punto di FM o LLM.
- NetApp SnapLock* offre la funzionalità di disco immutabile sui sistemi di storage basati su ONTAP per il controllo delle versioni dei set di dati. L'architettura microcore è progettata per proteggere i dati dei clienti con il motore FPolicy Zero Trust. NetApp garantisce la disponibilità dei dati dei clienti resistendo agli attacchi DoS (denial-of-service) quando un aggressore interagisce con un LLM in un modo che richiede un consumo particolarmente elevato di risorse.
- NetApp Cloud Data Sense* aiuta a identificare, mappare e classificare le informazioni personali presenti nei set di dati aziendali, a emanare policy, a soddisfare i requisiti di privacy in sede o nel cloud, a migliorare la sicurezza e a rispettare le normative.
- Classificazione NetApp BlueXP*, basata su Cloud Data Sense. I clienti possono automaticamente scansionare, analizzare, categorizzare e agire sui dati in tutto il patrimonio di dati, rilevare rischi per la sicurezza, ottimizzare l'archiviazione e accelerare le distribuzioni cloud. Combina servizi di archiviazione e dati tramite il suo piano di controllo unificato. I clienti possono utilizzare istanze GPU per l'elaborazione e ambienti multicloud ibridi per la suddivisione in livelli di archiviazione a freddo e per archivi e backup.
- Dualità file-oggetto NetApp *. NetApp ONTAP consente l'accesso a doppio protocollo per NFS e S3. Con questa soluzione, i clienti possono accedere ai dati NFS dai notebook Amazon AWS SageMaker tramite bucket S3 da NetApp Cloud Volumes ONTAP. Ciò offre flessibilità ai clienti che necessitano di un facile accesso a fonti di dati eterogenee con la possibilità di condividere dati sia da NFS che da S3. Ad esempio, per ottimizzare FM come i modelli di generazione di testo Llama 2 di Meta su SageMaker con accesso a bucket di file-oggetti.

Il servizio * NetApp Cloud Sync* offre un modo semplice e sicuro per migrare i dati verso qualsiasi destinazione, nel cloud o in locale. Cloud Sync trasferisce e sincronizza in modo fluido i dati tra storage locali o cloud, NAS e archivi di oggetti.

- NetApp XCP* è un software client che consente migrazioni di dati da qualsiasi dispositivo a NetApp e da NetApp a NetApp in modo rapido e affidabile. XCP offre inoltre la possibilità di spostare in modo efficiente grandi quantità di dati dai file system Hadoop HDFS a ONTAP NFS, S3 o StorageGRID , mentre l'analisi dei file XCP fornisce visibilità nel file system.
- NetApp DataOps Toolkit* è una libreria Python che semplifica per data scientist, DevOps e data engineer l'esecuzione di varie attività di gestione dei dati, come il provisioning, la clonazione o lo snapshot quasi istantaneo di un volume di dati o di un'area di lavoro JupyterLab, supportati da storage NetApp scalabile ad alte prestazioni.

Sicurezza dei prodotti NetApp. Gli LLM potrebbero rivelare inavvertitamente dati riservati nelle loro risposte, il che rappresenta una preoccupazione per i CISO che studiano le vulnerabilità associate alle applicazioni di intelligenza artificiale che sfruttano gli LLM. Come sottolineato da OWASP (Open Worldwide Application Security Project), problemi di sicurezza quali l'avvelenamento dei dati, la perdita di dati, il diniego di servizio e le iniezioni rapide all'interno degli LLM possono avere ripercussioni sulle aziende, in quanto potrebbero esporre i dati ad accessi non autorizzati da parte degli aggressori. I requisiti di archiviazione dei dati dovrebbero includere controlli di integrità e snapshot immutabili per dati strutturati, semi-strutturati e non strutturati. Per il controllo delle versioni dei set di dati vengono utilizzati NetApp Snapshots e SnapLock . Offre un rigoroso controllo degli accessi basato sui ruoli (RBAC), nonché protocolli sicuri e crittografia standard del settore per proteggere sia i dati inattivi che quelli in transito. Cloud Insights e Cloud Data Sense insieme

offrono funzionalità che ti aiutano a identificare in modo forense la fonte della minaccia e a stabilire le priorità dei dati da ripristinare.

* ONTAP AI con DGX BasePOD*

L'architettura di riferimento NetApp ONTAP AI con NVIDIA DGX BasePOD è un'architettura scalabile per carichi di lavoro di machine learning (ML) e intelligenza artificiale (AI). Per la fase di formazione critica degli LLM, i dati vengono solitamente copiati dall'archivio dati al cluster di formazione a intervalli regolari. I server utilizzati in questa fase sfruttano le GPU per parallelizzare i calcoli, creando un'enorme richiesta di dati. Soddisfare le esigenze di larghezza di banda I/O grezza è fondamentale per mantenere un elevato utilizzo della GPU.

* ONTAP AI con NVIDIA AI Enterprise*

NVIDIA AI Enterprise è una suite end-to-end e cloud-native di software di analisi dei dati e intelligenza artificiale, ottimizzata, certificata e supportata da NVIDIA per l'esecuzione su VMware vSphere con sistemi certificati NVIDIA. Questo software semplifica e velocizza l'implementazione, la gestione e il ridimensionamento dei carichi di lavoro di intelligenza artificiale nel moderno ambiente cloud ibrido. NVIDIA AI Enterprise, basato su NetApp e VMware, offre una gestione dei dati e dei carichi di lavoro di intelligenza artificiale di livello aziendale in un pacchetto semplificato e familiare.

Piattaforme cloud 1P

Le offerte di archiviazione cloud completamente gestite sono disponibili in modo nativo su Microsoft Azure come Azure NetApp Files (ANF), su AWS come Amazon FSx for NetApp ONTAP (FSx ONTAP) e su Google come Google Cloud NetApp Volumes (GNCV). 1P è un file system gestito ad alte prestazioni che consente ai clienti di eseguire carichi di lavoro di intelligenza artificiale ad alta disponibilità con maggiore sicurezza dei dati nei cloud pubblici, per ottimizzare LLM/FM con piattaforme ML native del cloud come AWS SageMaker, Azure-OpenAI Services e Vertex AI di Google.

Suite di soluzioni per i partner NetApp

Oltre ai suoi principali prodotti, tecnologie e capacità di dati, NetApp collabora a stretto contatto anche con una solida rete di partner di intelligenza artificiale per offrire valore aggiunto ai clienti.

- I NVIDIA Guardrails* nei sistemi di intelligenza artificiale servono come misure di salvaguardia per garantire l'uso etico e responsabile delle tecnologie di intelligenza artificiale. Gli sviluppatori di intelligenza artificiale possono scegliere di definire il comportamento delle applicazioni basate su LLM su argomenti specifici e impedire loro di impegnarsi in discussioni su argomenti indesiderati. Guardrails, un toolkit open source, offre la possibilità di connettere un LLM ad altri servizi in modo semplice e sicuro, per creare sistemi conversazionali LLM affidabili e sicuri.

Domino Data Lab fornisce strumenti versatili e di livello aziendale per la creazione e la produzione di intelligenza artificiale generativa: veloce, sicura ed economica, indipendentemente da dove ti trovi nel tuo percorso verso l'intelligenza artificiale. Con la piattaforma Enterprise MLOps di Domino, gli scienziati dei dati possono utilizzare gli strumenti preferiti e tutti i loro dati, addestrare e distribuire modelli facilmente ovunque e gestire rischi e costi in modo efficace, il tutto da un unico centro di controllo.

Modzy per Edge AI. NetApp e Modzy hanno stretto una partnership per fornire l'intelligenza artificiale su larga scala a qualsiasi tipo di dati, tra cui immagini, audio, testo e tabelle. Modzy è una piattaforma MLOps per la distribuzione, l'integrazione e l'esecuzione di modelli di intelligenza artificiale, che offre agli scienziati dei dati le capacità di monitoraggio dei modelli, rilevamento delle derive e spiegabilità, con una soluzione integrata per un'inferenza LLM senza interruzioni.

Run:AI e NetApp hanno stretto una partnership per dimostrare le capacità uniche della soluzione NetApp ONTAP AI con la piattaforma di gestione dei cluster Run:AI, al fine di semplificare l'orchestrazione dei carichi di lavoro di intelligenza artificiale. Divide e unisce automaticamente le risorse GPU, ed è progettato per scalare le pipeline di elaborazione dati su centinaia di macchine con framework di integrazione integrati per Spark, Ray, Dask e Rapids.

Conclusione

L'intelligenza artificiale generativa può produrre risultati efficaci solo quando il modello viene addestrato su grandi quantità di dati di qualità. Sebbene gli LLM abbiano raggiunto traguardi notevoli, è fondamentale riconoscerne i limiti, le sfide progettuali e i rischi associati alla mobilità e alla qualità dei dati. Gli LLM si basano su set di dati di formazione ampi e eterogenei provenienti da fonti di dati eterogenee. Risultati imprecisi o distorti generati dai modelli possono mettere a repentaglio sia le aziende che i consumatori. Questi rischi possono corrispondere a vincoli per gli LLM che emergono potenzialmente dalle sfide di gestione dei dati associate alla qualità dei dati, alla sicurezza dei dati e alla mobilità dei dati. NetApp aiuta le organizzazioni a far fronte alle complessità create dalla rapida crescita dei dati, dalla mobilità dei dati, dalla gestione multi-cloud e dall'adozione dell'intelligenza artificiale. Su larga scala, l'infrastruttura di intelligenza artificiale e la gestione efficiente dei dati sono fondamentali per definire il successo delle applicazioni di intelligenza artificiale come l'intelligenza artificiale generativa. È fondamentale che i clienti coprano tutti gli scenari di implementazione senza compromettere la capacità di espandersi in base alle esigenze aziendali, mantenendo al contempo il controllo sull'efficienza dei costi, sulla governance dei dati e sulle pratiche etiche di intelligenza artificiale. NetApp lavora costantemente per aiutare i clienti a semplificare e accelerare le implementazioni dell'intelligenza artificiale.