DELL Technologies

AMD
together we advance_

2024 | White paper

Potenziare le aziende con l'AI: l'ingresso nell'era della scelta

Sommario

DELL Technologies

AMD
together we advance_

2024 | White paper

L'opportunità di trasformare i settori economici con l'AI

Oggi non esiste opportunità migliore per trasformare il business preparandolo al futuro dell'innovazione grazie all'AI. I dati raccolti nel report Accenture Vision Technology 2023 indicano che il 98% dei dirigenti globali concorda sul fatto che i foundation model di AI svolgeranno un ruolo importante nelle strategie della loro organizzazione nei prossimi 3-5 anni.[1]

L'AI è diventata incredibilmente utile per le aziende che operano in settori quali la vendita al dettaglio, la sanità e i servizi finanziari, grazie alla sua capacità di potenziare l'efficienza delle attività, promuovere l'innovazione e migliorare i processi decisionali. Tuttavia, nonostante i vantaggi, quando si tratta di integrare l'AI si avverte ancora una barriera all'ingresso a causa di alcune errate convinzioni comuni.

È necessario un team di sviluppatori AI per iniziare:

Sebbene siano comunque preziose per sviluppare soluzioni di AI avanzate e comprenderne i principi sottostanti, le competenze in materia di Data Science non sono più un prerequisito. Vi è stata infatti una proliferazione di strumenti di AI intuitivi, piattaforme come Hugging Face e modelli specifici per singole attività che rimuovono gran parte della complessità implicita nello sviluppo delle soluzioni di AI.

Per ottenere risultati, è necessario spendere decine di milioni in hardware:

Questa idea errata mina gravemente la diversità delle risorse di AI oggi disponibili. Sebbene queste risorse comunemente note siano spesso potenti e ben supportate, potrebbero non sempre essere la scelta più idonea o conveniente per ogni azienda.

È necessario lavorare instancabilmente per acquisire gli acceleratori:

Anche se gli acceleratori sono eccellenti per i carichi di lavoro AI intensivi, le aziende potrebbero non aver bisogno di così tanta potenza di elaborazione per le proprie applicazioni di AI. Inoltre, aspettare un periodo di tempo eccessivamente lungo per avere accesso agli acceleratori leader del mercato semplicemente non è realistico. In molti casi, le CPU ottimizzate per l'AI possono effettivamente erogare le prestazioni e l'efficienza necessarie per produrre analisi e decisioni assistite dall'AI in tempo reale e sono una soluzione molto più conveniente e adattabile.

[1] Accenture, 30 marzo 2023, "Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds", https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds

Fortunatamente, il panorama dell'AI è in evoluzione. Insieme, **Dell** e **AMD** stanno collaborando per sfatare questi miti rendendo le tecnologie e gli strumenti di AI accessibili a una gamma più ampia di utenti con un'infrastruttura end-to-end progettata per supportare le esigenze dell'AI di oggi.

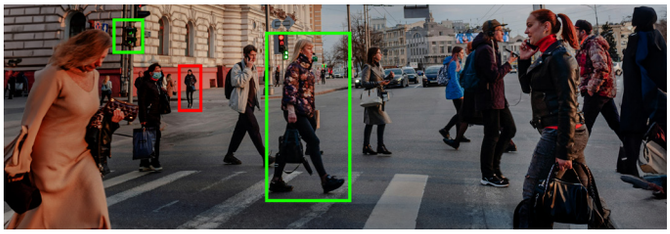
È possibile iniziare con un modello già ottimizzato, uno stack software affidabile e un sistema hardware versatile, tutti elementi disponibili apertamente grazie alla partnership di Dell e AMD. Poter disporre di acceleratori che sono sempre più limitati, di un gruppo consistente di ingegneri AI qualificati o di risorse per l'implementazione di cluster di grandi dimensioni nel cloud non è più un requisito per sfruttare l'AI.

Quel che la collaborazione tra **Dell** e **AMD** offre è un ecosistema unificato di hardware e software, progettato per consentire agli sviluppatori di creare soluzioni di AI end-to-end che integrano apprendimento induttivo, ottimizzazione e inferenza in modo semplice ed efficiente. Con il supporto di **Hugging Face**, disponiamo ora di un portafoglio sempre più ampio di modelli che vengono eseguiti su server Dell PowerEdge dotati di processori AMD EPYC™ o acceleratori AMD Instinct™ MI300X, affinché gli sviluppatori possano eseguire operazioni di ottimizzazione, applicazione dell'apprendimento induttivo e implementazione per inferenza. Gli investimenti in AMD ROCm™ e AMD ZenDNN™, nonché le partnership con i framework PyTorch, Tensorflow e ONNX Runtime, sono enabler fondamentali degli sviluppatori di AI applicata che sperimentano la democratizzazione dell'AI. Il diagramma dello stack riportato di seguito illustra in dettaglio i componenti che costituiscono l'ecosistema AI unificato di Dell e AMD.

L'AI nei vari settori

Con la diversificazione delle risorse e l'enfasi posta sull'innovazione open source, l'AI sta inserendosi in molti settori, tra cui assistenza clienti, finanza e banche, sanità e vendita al dettaglio, solo per citarne alcuni. In questi settori, tuttavia, l'AI consente globalmente alle organizzazioni di sbloccare il potenziale dei loro dati proprietari e di reinventare i flussi di lavoro di AI grazie alle funzionalità chiave di analisi dei dati, automazione, personalizzazione e analisi predittiva. Le librerie AMD ROCm e ZenDNN accelerano inoltre questi flussi di lavoro di AI per fornire risultati quasi in tempo reale.

Analizziamo di seguito in che modo l'AI influenza vari settori.

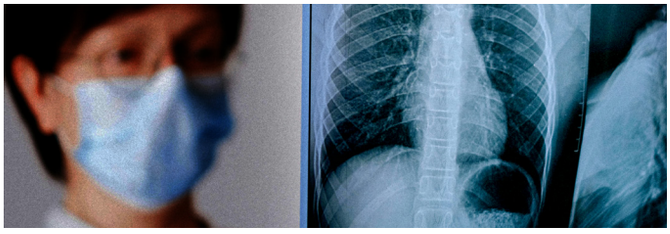


Settore automobilistico

L'AI viene utilizzata per il rilevamento di oggetti, il posizionamento in corsia e il processo decisionale nei veicoli autonomi. L'AI è inoltre in grado di prevedere quando un componente del veicolo potrebbe guastarsi, consentendo una manutenzione proattiva e riducendo i tempi di fermo.

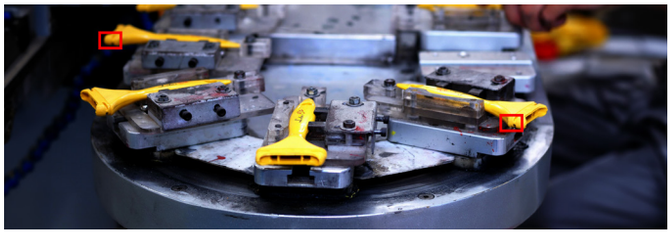
Vendita al dettaglio

L'AI può analizzare il comportamento dei clienti per fornire consigli personalizzati sui prodotti, migliorando quindi il coinvolgimento dei clienti e le vendite. Può inoltre ottimizzare i livelli delle scorte prevedendo la domanda e riducendo al minimo l'accumulo di scorte in eccesso o il loro esaurimento.



Medicali

L'AI può essere utilizzata nel settore sanitario per una varietà di applicazioni, tra cui analisi delle immagini medicali, diagnosi delle malattie, piani terapeutici personalizzati e scoperta di farmaci, con conseguente miglioramento degli esiti dei pazienti e riduzione dei costi.



Settore manifatturiero e industriale

L'AI può essere utilizzata nel settore manifatturiero e industriale per la manutenzione predittiva, il controllo qualità, l'ottimizzazione dei processi e il Supply Chain Management, con conseguente miglioramento dell'efficienza e riduzione dei downtime.

Servizi finanziari

L'AI può essere utilizzata nel settore finanziario e bancario per il rilevamento delle frodi, l'assessment dei rischi, l'assistenza clienti e l'analisi degli investimenti, determinando un miglioramento della sicurezza e un processo decisionale più informato.

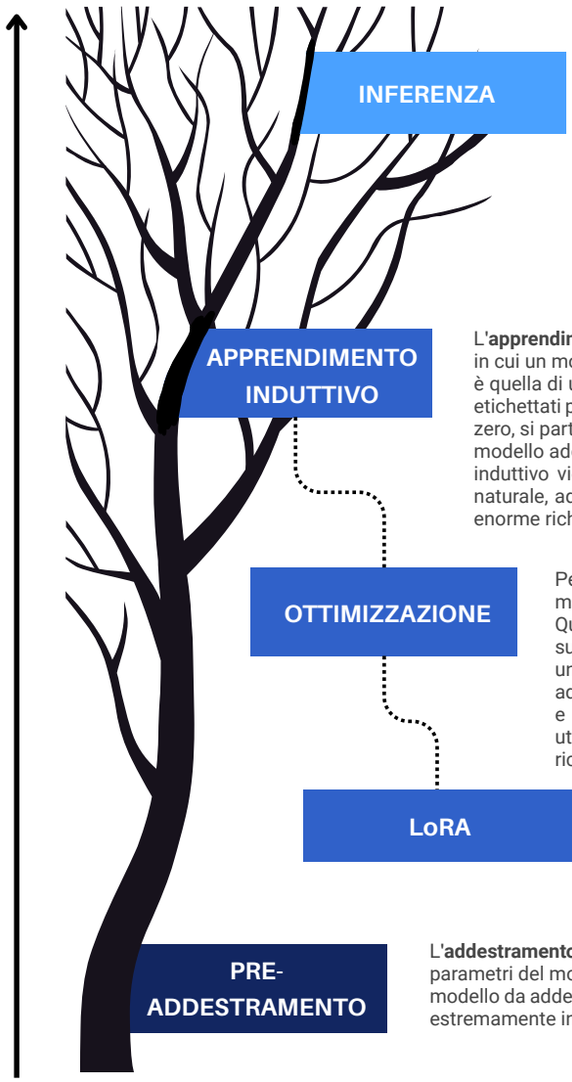
Automazione dei servizi

I chatbot basati sull'AI possono gestire le richieste dei clienti e fornire supporto, riducendo la necessità di intervento umano. L'AI può inoltre automatizzare attività ripetitive come l'inserimento di dati o l'elaborazione di documenti, migliorando l'efficienza e riducendo gli errori.

Che cosa i responsabili delle decisioni IT devono considerare

PER INIZIARE: ANALISI DEI PROCESSI DELL'AI

Prima di esaminare questi casi d'uso, esaminiamo più dettagliatamente il ciclo di vita dell'AI. Per ciclo di vita dell'AI (intelligenza artificiale) si intendono le fasi coinvolte nello sviluppo, nell'implementazione e nella manutenzione di un sistema di AI. Sebbene le metodologie specifiche e la terminologia possano variare, un ciclo di vita dell'AI tipico include sempre l'addestramento e l'inferenza del modello.



L'**inferenza** è il processo di utilizzo di un modello di AI addestrato per eseguire previsioni, classificazioni o decisioni basate su dati nuovi e invisibili. In questa fase del ciclo di vita dell'AI, il modello addestrato viene implementato e utilizzato in un ambiente di produzione per eseguire l'attività specifica per cui è stato progettato. Durante l'inferenza, il modello acquisisce i dati di input, li elabora utilizzando i pattern e le relazioni che ha appreso nella fase di addestramento e produce un output, ad esempio una previsione, una classificazione o un testo nei carichi di lavoro di GenAI. Questo output può quindi essere utilizzato per varie applicazioni, ad esempio per consigliare prodotti a un utente, rilevare anomalie in un sistema o riconoscere oggetti in un'immagine. Nell'elaborazione del linguaggio naturale, la RAG (Retrieval-Augmented Generation) migliora l'accuratezza durante l'implementazione dei modelli, consentendo ad essi di accedere e sfruttare fonti di conoscenza esterne e quindi di comprendere e rispondere meglio agli input in linguaggio naturale in una più vasta gamma di contesti.

L'**apprendimento induttivo**, noto anche come ri-addestramento, è una tecnica di apprendimento automatico in cui un modello addestrato per un'attività viene riutilizzato per una seconda attività correlata. L'idea generale è quella di utilizzare la conoscenza che un modello ha appreso da un'attività con molti dati di addestramento etichettati per una nuova attività che non dispone di molti dati. Anziché iniziare il processo di apprendimento da zero, si parte dai pattern appresi dalla risoluzione di un'attività correlata, in pratica aggiornando i parametri del modello addestrato per la nuova attività correlata in modo che si adattino al meglio ad essa. L'apprendimento induttivo viene utilizzato principalmente nelle attività di visione artificiale e di elaborazione del linguaggio naturale, ad esempio l'analisi del sentiment, per ridurre la quantità di potenza di elaborazione generalmente enorme richiesta per queste attività.

Per **ottimizzazione** nel ciclo di vita dell'AI si intende il processo di ulteriore addestramento di un modello di apprendimento automatico pre-addestrato su un nuovo dataset o una nuova attività. Questa tecnica è comunemente utilizzata nell'apprendimento induttivo, in cui un modello addestrato su un dataset di grandi dimensioni per un'attività specifica viene adattato a un'attività correlata o a un dataset più piccolo. L'ottimizzazione consente il trasferimento delle conoscenze dal modello pre-addestrato alla nuova attività o al nuovo dataset, spesso determinando una convergenza più rapida e prestazioni migliori rispetto all'addestramento di un modello partendo da zero. È ampiamente utilizzata in varie applicazioni di AI, tra cui visione artificiale, elaborazione del linguaggio naturale e riconoscimento vocale.

LoRA (Low Rank Adaptation) è una nuova tecnica per l'ottimizzazione di modelli di apprendimento approfondito che riducono il numero di parametri addestrabili e che permettono l'efficiente passaggio da un'attività all'altra. In questo modo il processo di ottimizzazione richiede una minore intensità di elaborazione ed è più accessibile con hardware largamente disponibile.

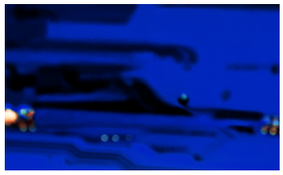
L'**addestramento** comporta la selezione dell'architettura appropriata per il modello o l'algoritmo, l'ottimizzazione dei parametri del modello e la valutazione delle prestazioni del modello internamente. A seconda della dimensione del modello da addestrare, l'aggiornamento dei parametri in ogni iterazione dell'addestramento può essere un processo estremamente intensivo in termini di dati ed elaborazione.

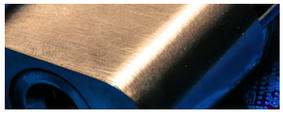
Figura 1. Il ciclo di vita dell'AI

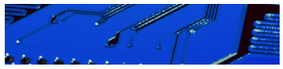
SCELTE CRITICHE

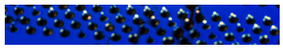
| Prestazioni

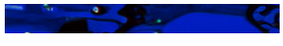
In molte di queste applicazioni concrete, l'adozione di decisioni in tempo reale o quasi reale è un fattore cruciale per il successo. Ad esempio, le attività fraudolente nelle transazioni finanziarie o nelle richieste di indennizzo assicurativo devono essere identificate tempestivamente per impedire perdite finanziarie e proteggere le risorse aziendali. In uno scenario di produzione, eventuali difetti nella linea di assemblaggio o nelle condizioni di fabbrica devono essere monitorati in modo dinamico per garantire la qualità. In pratica, il processore che gestisce il carico di lavoro di inferenza deve essere ottimizzato per elaborare i flussi di dati in ingresso in modo rapido ed efficiente. I server Dell PowerEdge abbinati ai processori AMD EPYC formano una combinazione versatile, particolarmente adatta per la gestione dei carichi di lavoro di inferenza sull'edge e per le attività che implicano High Performance Computing, cloud computing e analisi dei Big Data.











| Sicurezza dei dati

La **sicurezza dei dati** è fondamentale per il successo dei sistemi di AI, in particolare di quelli che sfruttano l'AI generativa, ed è un aspetto di grande importanza per i leader tecnologici che puntano a incorporare l'AI nelle loro operazioni. I sistemi di AI si basano in genere su enormi quantità di dati, che possono includere informazioni riservate e sensibili come dati personali, finanziari o proprietari. La salvaguardia di questi dati è fondamentale per impedirne l'accesso non autorizzato o il furto, oltre che per garantire la precisione, l'affidabilità e la coerenza dei modelli di AI e delle previsioni.

L'**elaborazione riservata** è una tecnologia che facilita l'elaborazione dei dati in un'enclave sicura, proteggendoli dall'accesso non autorizzato o da manipolazioni da parte di soggetti non autorizzati, ad esempio il provider di cloud e altri utenti.[2] Per isolare i dati durante l'elaborazione, vengono utilizzate la crittografia e altre misure di sicurezza. AMD Infinity Guard, una raccolta di sofisticate funzionalità di protezione integrate nei processori AMD EPYC, supporta l'elaborazione riservata mediante l'impiego della tecnologia SEV (Secure Encrypted Virtualization), che crittografa le macchine virtuali utilizzando una chiave nota solo al processore. L'obiettivo di questi servizi è fornire ambienti di esecuzione affidabili basati su hardware con AMD SEV-SNP (SEV-Secure Nested Paging), che migliora le protezioni guest per contribuire alla difesa dalle minacce esterne.

L'**apprendimento federato** è un altro metodo per preservare la sicurezza dei dati, che opera addestrando un modello centrale attraverso dispositivi o server decentralizzati.[3] Anziché trasferire tutti i dati in una posizione centrale, ogni dispositivo addestra il modello in locale e vengono condivisi solo gli aggiornamenti del modello. Questo approccio preserva la privacy e consente l'apprendimento collaborativo senza condivisione dei dati non elaborati. La piattaforma di AI federata di Dell Technologies consente di eseguire processi computazionali, algoritmi di AI e ML su dataset sull'edge della rete durante la raccolta, condividendo in rete solo modelli matematici, metadati e risultati delle query con altri dispositivi edge, i data center o il cloud. Questo scambio migliora i risultati consentendo l'estrazione quasi in tempo reale di informazioni utili da dataset distribuiti di grandi dimensioni senza rivelare i dati ed eventuali proprietà intellettuali.

[2] Advanced Micro Devices, Inc., 30 agosto 2023, "MD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings", https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html
Advanced Micro Devices, Inc., 2021, Solution brief "Data Center Solutions, Confidential Computing", https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

[3] Analytics Vidhya, dicembre 2023, "Federated Learning: A Beginner's Guide", https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data
Dell Technologies, 2021, Solution brief "A federated learning platform for real-time artificial intelligence", https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

DIMENSIONAMENTO DELLA SOLUZIONE

| Equilibrio tra costi e innovazione

Individuare il giusto equilibrio tra costi e innovazione garantisce che le soluzioni di AI siano non solo finanziariamente fattibili, ma anche di grande impatto, generando un valore reale sia per le aziende che per gli utenti. Un fattore chiave per trovare questo equilibrio sta nell'identificare un hardware che risolva i casi d'uso e che si integri facilmente nell'infrastruttura esistente. Nel moderno mercato dell'hardware per l'AI, contribuiscono alla scarsità di acceleratori l'aumento della loro domanda in vari settori, nonché i vincoli di capacità produttiva, le sfide della logistica e la carenza di semiconduttori.

Tuttavia, le CPU sono già un componente standard nella maggior parte dei data center, il che rende l'integrazione più semplice e a costi contenuti rispetto all'aggiunta di hardware di accelerazione completamente nuovo. Le CPU ottimizzate per l'AI possono sfruttare il software e gli strumenti esistenti, riducendo la necessità di vaste attività di retooling o riaddestramento. Le CPU offrono inoltre maggiore flessibilità ed efficienza per un'ampia gamma di attività oltre all'AI, consentendo un uso più versatile delle risorse all'interno del data center. L'aggiornamento del data center con i server Dell PowerEdge dotati di processori AMD EPYC supporta l'esecuzione dei carichi di lavoro esistenti e al tempo stesso li rende pronti per progredire verso maggiori innovazioni ed efficienze basate sull'AI.

| Semplicità e flessibilità

La semplicità e la flessibilità del sistema di AI sono essenziali per creare soluzioni di AI efficaci, adattabili e scalabili nel lungo periodo. Avere accesso a una suite di framework e ottimizzazioni software a integrazione dell'hardware migliora le prestazioni senza dover dedicare ulteriore tempo e sforzo all'integrazione multipiattaforma. Queste qualità sono di particolare importanza quando si tratta di affrontare i carichi di lavoro di AI misti, che implicano una combinazione di diversi tipi di attività di AI come l'addestramento, l'inferenza e l'elaborazione dei dati.

AMD e Dell Technologies gestiscono i carichi di lavoro di AI misti attraverso una combinazione di soluzioni hardware e software. I processori AMD EPYC offrono potenza per l'High Performance Computing, con funzioni come il multithreading simultaneo (SMT), nonché un elevato numero di core, consentendo un'efficiente elaborazione parallela per i carichi di lavoro di AI. Questi processori sono ottimizzati per le attività di AI e offrono prestazioni elevate sia per i carichi di lavoro di addestramento che per quelli di inferenza. I server Dell PowerEdge, dotati di processori AMD EPYC, forniscono una piattaforma scalabile e flessibile per l'implementazione dei carichi di lavoro di AI. Inoltre, la suite Dell OpenManage Software offre strumenti di gestione per ottimizzare l'allocazione delle risorse e il monitoraggio delle prestazioni per carichi di lavoro di AI misti.

AMD offre anche UIF (Unified Inference Frontend), che sfrutta le versioni a prestazioni avanzate di ciascuno degli stack software odierni e che utilizza la libreria AMD ZenDNN per i processori AMD EPYC, lo stack open source AMD ROCm per gli acceleratori AMD Instinct, nonché uno stack software per i SoC adattivi AMD. AMD ROCm è inoltre progettato per operare con un'ampia gamma di CPU e acceleratori AMD, tra cui prodotti di livello professionale e consumer.

| Garanzia di spiegabilità

L'**AI spiegabile** svolge un ruolo fondamentale nel garantire trasparenza, affidabilità ed efficacia nelle applicazioni di intelligenza artificiale. L'AI spiegabile fornisce informazioni approfondite sul modo in cui i modelli di AI prendono decisioni, facendo luce sui fattori sottostanti e sui processi di ragionamento. Questa trasparenza è fondamentale per ottenere la fiducia delle entità interessate, soprattutto in settori sensibili come quello sanitario, finanziario e della giustizia penale, dove le decisioni hanno un impatto diretto sulla vita delle persone.

Human-in-the-loop I sistemi AI sfruttano l'intelligenza umana per migliorare le prestazioni dell'AI e mitigare i pregiudizi algoritmici. Integrando la supervisione umana, questi sistemi sono in grado di gestire situazioni complesse e ambigue in modo più efficace, garantendo che le soluzioni di AI siano in linea con le norme etiche e sociali. Inoltre, il coinvolgimento umano consente il perfezionamento e l'adattamento continui dei modelli di AI in base al feedback reale, favorendo il miglioramento iterativo e l'affidabilità a lungo termine. Questi approcci sono essenziali per creare sistemi di AI responsabili e inclusivi che servano al meglio gli interessi della società.

Scenari del mondo reale

Scalers AI ha collaborato con Dell e AMD per dimostrare le funzionalità dei server Dell PowerEdge dotati di processori AMD. Ecco come queste tecnologie vengono sfruttate per l'addestramento, l'apprendimento induttivo e l'inferenza in scenari di vendita al dettaglio e del settore sanitario.

VENDITA AL DETTAGLIO

Scalers AI ha creato Retail Inventory Management Reference Solution, un sistema progettato per monitorare e gestire i livelli delle scorte sugli scaffali dei punti vendita al dettaglio attraverso l'implementazione di un modello di AI per il rilevamento degli oggetti. Questa soluzione di riferimento sfrutta il modello SSD_MobileNet_V2 per identificare e riconoscere i prodotti sugli scaffali dei punti vendita, consentendo il conteggio automatico dell'inventario e il monitoraggio esatto dei livelli delle scorte. Il modello è stato sottoposto ad apprendimento induttivo utilizzando il dataset di immagini SKU110K, comprendente 23.000 immagini di Roboflow. Sfruttando algoritmi di visione artificiale e apprendimento automatico, il sistema è in grado di rilevare quando gli articoli stanno esaurendosi o si sono esauriti, fornendo avvisi al personale del punto vendita per un tempestivo reintegro o ricostituzione delle scorte.

Questa soluzione utilizza il server Dell PowerEdge R7615 con il processore AMD EPYC 9354P a 32 core.

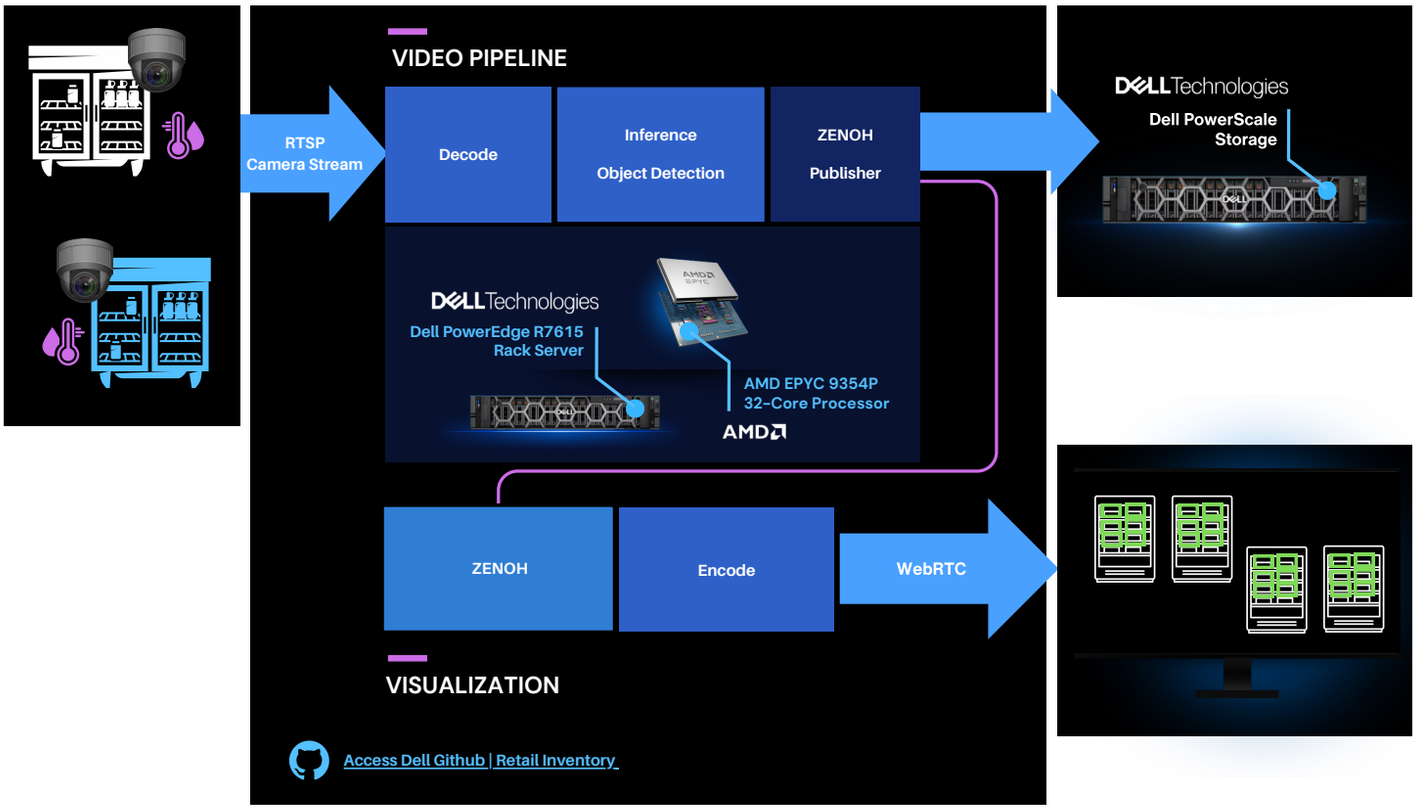


Figura 2. Diagramma dell'architettura della Retail Inventory Management Reference Solution

SETTORE SANITARIO

L'imaging medicale basato sull'AI offre un immenso valore grazie alla sua capacità di potenziare l'assistenza sanitaria migliorando l'accuratezza e l'efficienza della diagnostica e fornendo ai professionisti del settore sanitario informazioni approfondite e precise su condizioni che potrebbero essere difficili da rilevare a occhio nudo. Automatizzando l'analisi delle immagini medicali, l'AI riduce il tempo necessario per la diagnosi, consentendo decisioni terapeutiche più rapide e quindi migliorando l'esito dei pazienti.

Scalers AI ha sfruttato le funzionalità del server Dell PowerEdge R7625 con processori AMD EPYC 9554 a 64 core per creare una soluzione di imaging medicale basata sull'AI per il rilevamento della polmonite. Utilizzando algoritmi avanzati e tecniche di apprendimento automatico per analizzare immagini medicali come radiografie o scansioni TC, la soluzione favorisce maggiore velocità e accuratezza della diagnosi di polmonite nei pazienti. In questo modo si introduce un ulteriore livello di controllo assistito da computer, con la possibilità di aiutare gli operatori sanitari nella gestione di grandi volumi di dati di imaging in modo più efficiente.

Questa soluzione di riferimento utilizza il modello ResNet50 per analizzare le immagini radiografiche del torace ottenute dal dataset del NIH Clinical Center. Il suo obiettivo primario è rilevare la presenza o l'assenza di polmonite, eseguendo essenzialmente una classificazione binaria. Il modello è stato addestrato utilizzando il dataset Xray DICOM del dataset del NIH Clinical Center, utilizzando l'apprendimento induttivo con l'architettura ResNet50.

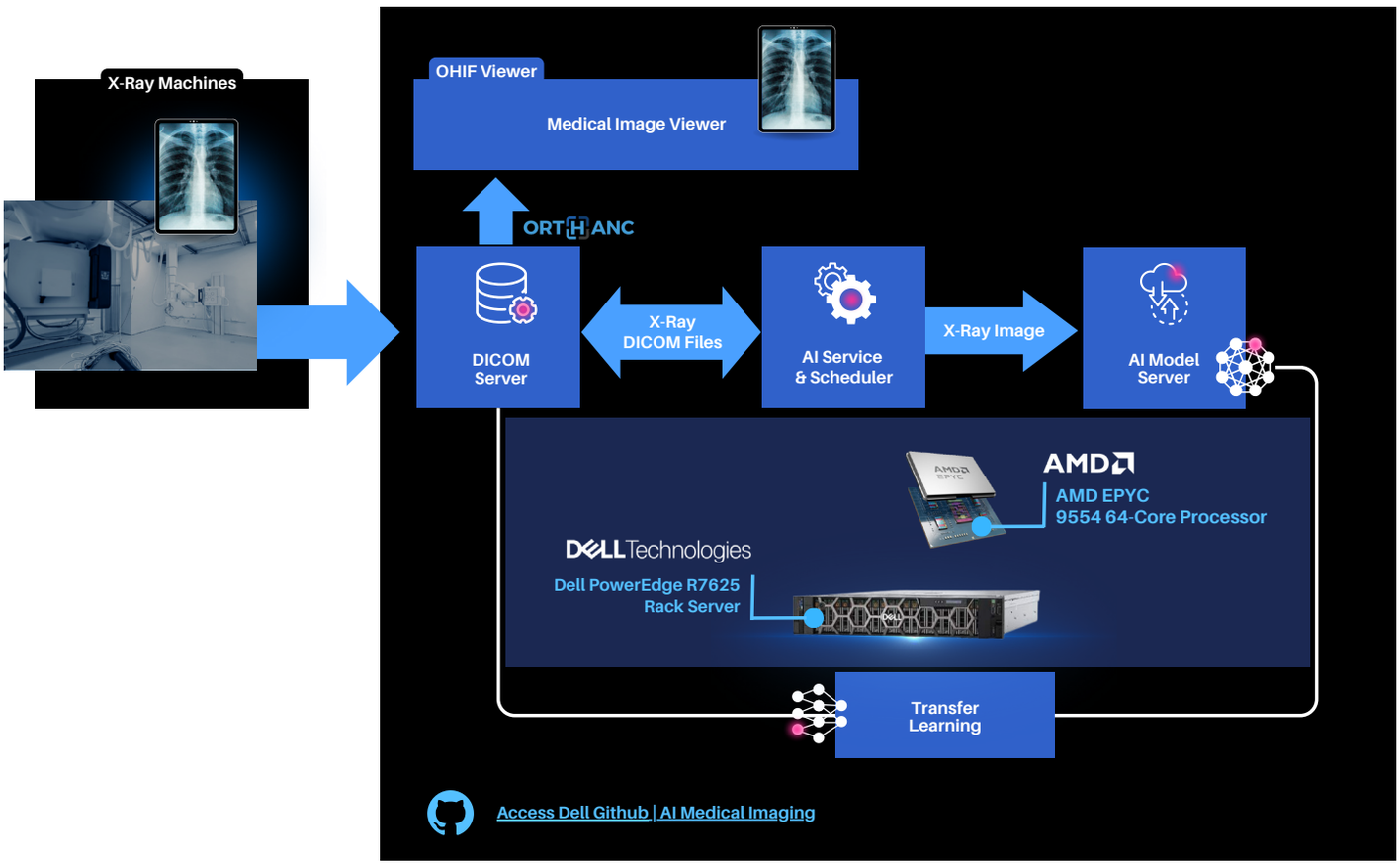
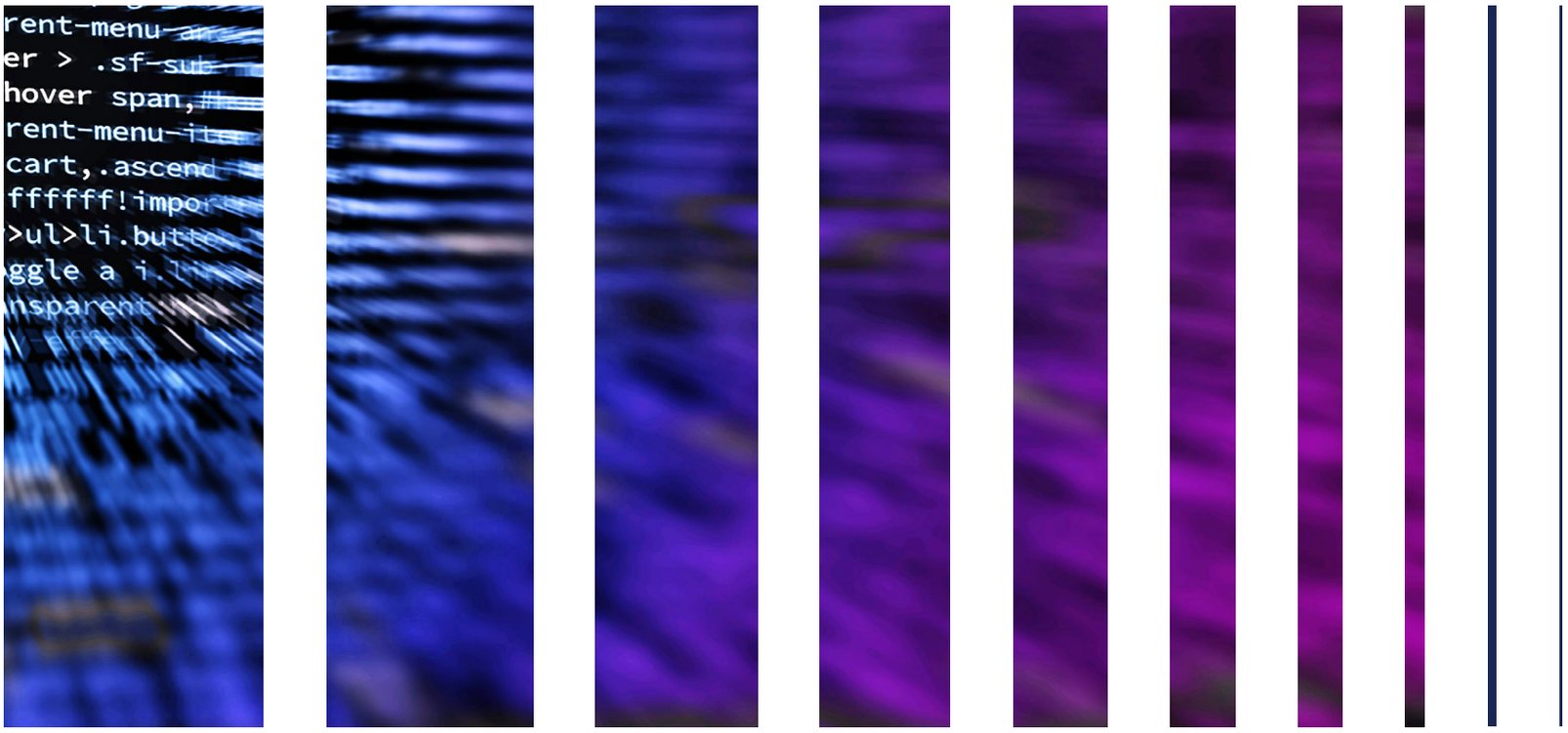


Figura 3. Diagramma dell'architettura della Medical AI Imaging Solution

Le nostre soluzioni

L'AI È PER TUTTI: DELL E AMD DEMOCRATIZZANO L'AI

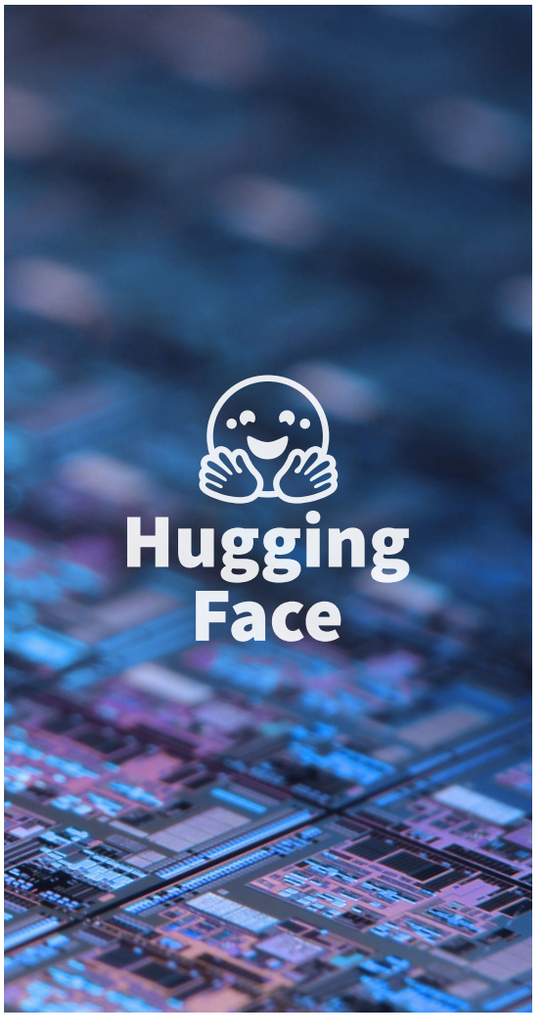
Questa collaborazione getta le basi per la democratizzazione dell'AI, essenziale per incentivare l'innovazione e promuovere l'inclusività nell'ecosistema AI. Dell e AMD stanno raggiungendo questo risultato consentendo a individui e organizzazioni di sfruttare l'AI e risolvere sfide specifiche nei rispettivi campi con una suite accessibile di potenti server dotati di CPU AMD e tecnologie di accelerazione all'avanguardia. I server Dell PowerEdge con acceleratori AMD Instinct MI300X sono in grado di gestire carichi di lavoro di AI di grandi dimensioni, come l'addestramento e l'ottimizzazione di LLM (Large Language Model), mentre i server Dell PowerEdge dotati di processori AMD EPYC sono eccezionali nella gestione dei carichi di lavoro di inferenza sull'edge. Oltre alla piattaforma hardware sottostante, AMD offre anche la libreria software ZenDNN per l'ottimizzazione dell'inferenza di apprendimento approfondito sulle CPU AMD, nonché la libreria software AMD ROCm per migliorare le capacità di addestramento, ottimizzazione e inferenza sugli acceleratori AMD Instinct. Tutte queste opzioni sono perfettamente collegate nell'UIF (Unified Inferencing Model) di AMD, attraverso il quale gli utenti possono creare soluzioni di AI end-to-end con flessibilità nella scelta dei framework software, delle ottimizzazioni software e della piattaforma hardware.

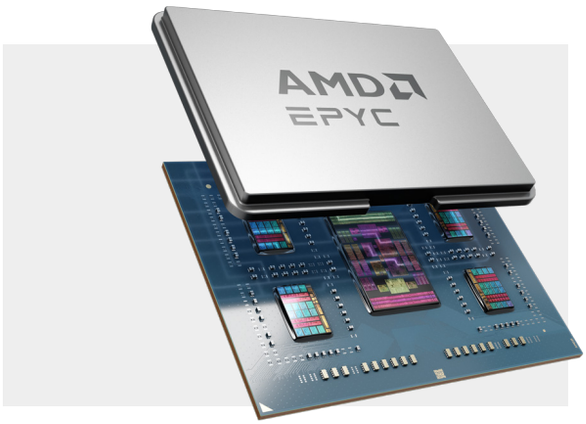


COLLABORAZIONE CON HUGGING FACE

Le aziende che intendono adottare l'AI possono iniziare sfruttando modelli preesistenti o flussi di lavoro di AI personalizzati per le loro esigenze specifiche direttamente offerti da Hugging Face, una piattaforma open source dedicata alla Data Science e all'apprendimento automatico. AMD ha avviato una collaborazione con Hugging Face, con l'obiettivo condiviso di erogare prestazioni di trasformatori di altissimo livello aggiungendo ottimizzazioni software specifiche di AMD a librerie e framework software che già si integrano perfettamente con le piattaforme AMD. Hugging Face sta collaborando attivamente con il team di progettazione di AMD per ottimizzare i modelli chiave in modo da ottenere massime prestazioni, incorporando AMD ROCm nella sua libreria di trasformatori e migliorando Optimum-AMD, una libreria specificamente progettata per le piattaforme AMD, al fine di agevolarne l'utilizzo da parte degli utenti di Hugging Face con modifiche minime al codice.

Inoltre, Dell Technologies ha recentemente unito le forze con Hugging Face per semplificare il processo di sviluppo, ottimizzazione e applicazione dei propri modelli di AI generativa (GenAI) open source utilizzando la community Hugging Face, il tutto su prodotti e servizi dell'infrastruttura Dell leader del settore. Sulla piattaforma Hugging Face è in fase di sviluppo un nuovo portale Dell, che includerà container e script personalizzati e dedicati per aiutare gli utenti a implementare in modo sicuro e senza sforzo i modelli open source disponibili su Hugging Face utilizzando i server e i sistemi di storage dei dati Dell. Le aziende possono ora sfruttare appieno le risorse di Hugging Face per implementare direttamente i modelli sui server Dell PowerEdge con processori AMD e creare soluzioni di AI end-to-end con i propri dati proprietari.



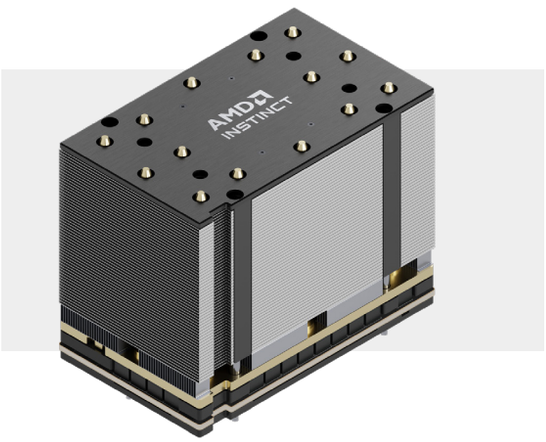


PROCESSORI AMD EPYC

AMD fornisce i progressi tecnologici necessari per i Modern Data Center basati su cloud tramite i suoi processori EPYC. Questi processori di tipo SoC (System On Chip) sono progettati da zero per rispondere in modo efficiente alle esigenze dei data center attuali e futuri. I processori AMD EPYC serie 9000 forniscono al data center fino a 128 core, 256 thread, 12 canali di memoria che supportano fino a 6 TB di memoria per socket e 128 corsie PCIe Gen5. A questo si aggiunge la soluzione per la sicurezza dei server x86 incorporata nell'hardware e all'avanguardia del settore. Integrando risorse essenziali di elaborazione, memoria, I/O e sicurezza nel SoC, i processori AMD EPYC offrono prestazioni di altissimo livello e agevolano la riduzione dei costi complessivi di gestione (TCO).

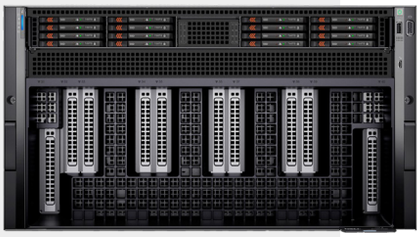
ACCELERATORI AMD INSTINCT MI300X

L'acceleratore AMD Instinct MI300X, basato sull'architettura all'avanguardia AMD CDNA 3, offre efficienza e prestazioni leader del settore per le applicazioni di AI e HPC più intensive. Dotato di 304 unità di elaborazione a prestazioni elevate, dispone di funzioni specifiche per l'AI, come il supporto per nuovi tipi di dati e la decodifica di foto e video, oltre a insuperabili 192 GB di memoria HBM3 su un singolo acceleratore.



PIATTAFORMA SOFTWARE OPEN SOURCE AMD ROCM 6

La piattaforma software open source AMD ROCm 6 è ottimizzata per massimizzare le prestazioni dei carichi di lavoro HPC (High Performance Computing) e AI degli acceleratori AMD Instinct MI300X. Estende inoltre il supporto per gli acceleratori AMD Instinct MI300X, assicurando la compatibilità con i framework software del settore. La piattaforma AMD ROCm racchiude una varietà di driver, strumenti di sviluppo e API che facilitano la programmazione degli acceleratori dal livello kernel alle applicazioni dell'utente finale e possono essere personalizzati per l'allineamento con requisiti specifici. È particolarmente adatta per applicazioni in ambito HPC (High Performance Computing), intelligenza artificiale (AI) ed elaborazione scientifica. Inoltre, la piattaforma AMD ROCm offre supporto per l'elaborazione a più acceleratori, incluso l'accesso RDMA (Remote Direct Memory Access) per la comunicazione server-nodo.



LINEA DI SERVER DELL POWEREDGE

L'investimento di Dell in AMD determina una scelta critica sul mercato per la democratizzazione dell'AI, come dimostrano le sue quattro piattaforme server con EPYC e il suo server rack di punta Dell PowerEdge XE9680 con acceleratori AMD Instinct MI300X. La più recente generazione di server Dell PowerEdge con processori AMD EPYC migliora sia l'agilità del business sia il time to market, con la possibilità di supportare carichi di lavoro di trasformazione quali database e analisi, virtualizzazione, software-defined storage, VDI (Virtual Desktop Infrastructure), containerizzazione, HPC, intelligenza artificiale e apprendimento automatico. I suoi server rack a un socket (CPU singola) offrono un equilibrio efficiente in termini di costo tra prestazioni e capacità di storage e sono progettati per crescere senza problemi con il business, mentre i server rack a due socket (CPU doppia) supportano i carichi di lavoro più impegnativi con un'ampia gamma di funzionalità.

Il server rack Dell PowerEdge XE9680 è una potente macchina di elaborazione dei dati specificamente su misura per le attività di AI. Grazie al supporto di otto acceleratori, è perfetto per i carichi di lavoro di apprendimento automatico (ML)/apprendimento approfondito (DL) e di inferenza, in particolare per quelli correlati all'addestramento dei LLM. Dotato di otto acceleratori MI300X, ciascuno con 192 GB di memoria HBM3 (High Bandwidth Memory 3) da 5,3 TB/s, per una capacità totale HBM3 di 1,5 TB per server e oltre 21 petaflops di prestazioni FP16, il server rack Dell PowerEdge XE9680 con acceleratori AMD Instinct MI300X è pronto per estendere ulteriormente l'accessibilità della GenAI alle aziende. Ciò consente di addestrare modelli più grandi, ridurre al minimo l'ingombro del data center, diminuire il TCO e ottenere un vantaggio competitivo.

Riepilogo

Il rapido ritmo dell'innovazione alimentato dall'AI sta rivoluzionando i carichi di lavoro dei data center più velocemente di qualsiasi altra trasformazione tecnologica. Per supportare questi progressi tecnologici, Dell e AMD stanno lavorando per creare un ecosistema AI più inclusivo, innovativo e sviluppato in modo etico che incoraggi gli sviluppatori di tutti i settori a collaborare su risorse open source e a promuovere l'innovazione della GenAI di oggi. Sia che i requisiti delle prestazioni della soluzione di AI in uso siano soddisfatti su processori AMD EPYC o su server con acceleratori AMD Instinct, offriamo la flessibilità necessaria per eseguire i carichi di lavoro di AI sulle nostre piattaforme hardware, consentendo di sfruttare in modo ottimale il meglio di ciò che Dell e AMD hanno da offrire.

RIFERIMENTI

Immagini AMD: AMD.com, AMD Partner Resource Library, **https://www.amd.com/en/partner/resources/resource-library.html**

Immagini Dell: **Dell.com**