Red Hat OpenShift AI

Una piattaforma open per i modelli di IA nel cloud ibrido

Punti chiave

Operatività e scalabilità dell'inferenza IA e dell'Agentic AI su una piattaforma applicativa collaudata.

Maggiore efficienza operativa dell'AI/ML grazie a un'esperienza utente omogenea che semplifica la collaborazione tra data scientist, data engineer, sviluppatori di applicazioni e team DevOps.

Maggiore flessibilità nel cloud ibrido perché la piattaforma permette di creare, addestrare, distribuire e monitorare i carichi di lavoro dell'AI/ML on premise, nel cloud e all'edge.

Adotta applicazioni intelligenti e IA generativa

Oggi, tecnologie quali intelligenza artificiale (IA), machine learning (ML) e deep learning (DL) influenzano profondamente le iniziative di modernizzazione delle applicazioni in moltissimi settori e attività. La necessità di innovare e di trarre valore strategico e nuove informazioni dai dati sta accelerando l'adozione di applicazioni cloud native basate sull'IA e delle metodologie MLOps e GenAIOps. In questo nuovo mondo, le sfide possono essere complesse: dall'aumento rapido dei costi dei modelli quando si passa alla produzione, alla personalizzazione complessa fino ai rigidi vincoli di deployment e alle operazioni necessarie per tenere il passo con l'innovazione. Le aziende hanno bisogno di soluzioni che riducano i costi di inferenza, semplifichino la scalabilità e il monitoraggio e si adattino ai cambiamenti continui.

Red Hat® AI accelera lo sviluppo e il deployment di soluzioni IA per le aziende in ambienti cloud ibridi. Si tratta di una piattaforma completa per la gestione dell'intero ciclo di vita dell'AI/ML, che offre funzionalità MLOps e GenAI Ops. Red Hat AI punta in particolare su quattro pilastri fondamentali:

- aumentare l'efficienza con un'inferenza rapida, flessibile ed efficiente;
- semplificare l'esperienza di connessione dei modelli ai dati;
- accelerare l'innovazione di Agentic AI; e
- garantire flessibilità e coerenza nella scalabilità dell'IA nel cloud ibrido.

Red Hat OpenShift® AI, basato su Red Hat OpenShift, una delle principali piattaforme applicative di cloud ibrido, è il prodotto di punta del portfolio Red Hat AI. Grazie alla base AI/ML con un potenziale elevato, la piattaforma di IA facilita ingegneri, data scientist e sviluppatori nella creazione e nel deployment in modo scalabile di modelli (generativi e predittivi) e applicazioni potenziate dall'IA. Adottando un'unica piattaforma condivisa, le aziende possono sperimentare nuovi strumenti, migliorare la collaborazione e accelerare i tempi di rilascio. Red Hat OpenShift AI coniuga l'ambiente self service tanto richiesto dai data scientist e dagli sviluppatori con la sicurezza richiesta dall'IT aziendale.

Accelera sviluppo, addestramento, test e deployment

Red Hat OpenShift AI è una piattaforma MLOps flessibile e scalabile basata su tecnologie open source, che offre funzionalità affidabili e coerenti dal punto di vista operativo per consentire ai team di sperimentare, distribuire modelli e rilasciare applicazioni innovative. OpenShift AI accelera la distribuzione delle applicazioni abilitate all'IA, aiutando le organizzazioni a passare dai primi progetti pilota a deployment solidi dal punto di vista operativo con maggiore velocità e controllo.

La piattaforma offre un'interfaccia utente (UI) integrata con strumenti per la creazione, l'addestramento, l'ottimizzazione, la distribuzione e il monitoraggio dei modelli di IA gen e predittiva. È possibile eseguire il deployment dei modelli in ambienti cloud ibridi, con un particolare accento sulla fornitura di un perimetro controllato e protetto per l'IA privata e sovrana. Questo approccio garantisce che i dati sensibili e i modelli di IA rimangano entro i limiti geografici o organizzativi designati, soddisfacendo severi requisiti normativi e di conformità.

Red Hat
OpenShift AI

Le previsioni degli analisti dell'IA gen

"Si prevede che l'IA sarà un fattore determinante per i budget destinati alle infrastrutture digitali nel 2026, poiché le organizzazioni stanno lavorando per adeguare il carico di lavoro e i requisiti dei dati alle scelte relative alle infrastrutture ibride. Il 90% dei responsabili delle decisioni ritiene che l'IA sarà un fattore determinante per il budget destinato alle infrastrutture digitali e per le scelte tecnologiche fino al 2026."[1]

Semplifica l'adozione dell'IA

Integrata con Red Hat OpenShift, OpenShift AI offre una piattaforma progettata per aumentare l'adozione dell'IA e rafforzare la fiducia nelle iniziative di IA, combinando comunità open source a un solido ecosistema di IA. In questo modo è possibile aumentare la flessibilità e la libertà di scelta della tecnologia AI/ML più adatta alla propria organizzazione. Gli utenti possono creare i propri modelli predittivi o iniziare con un modello di IA gen esterno, per poi perfezionarlo con la retrieval augmented generation (RAG) utilizzando uno dei numerosi server dei modelli forniti nella piattaforma. Quest'ultima offre un rapido accesso a modelli di terze parti ottimizzati e convalidati, come Llama, Mistral, DeepSeek e Granite, che vengono eseguiti in modo efficiente su vLLM e sono disponibili nel repository Red Hat AI su Hugging Face. Il catalogo consente agli utenti di esplorare questi modelli e aggiungerne di nuovi.

Migliora la coerenza operativa tra i team

Red Hat OpenShift AI offre un'esperienza utente coerente che consente a data scientist, ingegneri di IA, sviluppatori e team DevOps di collaborare in modo efficace per fornire soluzioni di IA tempestive. Offre accesso self service a flussi di lavoro collaborativi, accelerazione delle unità di elaborazione grafica (GPU) e operazioni semplificate, fornendo una distribuzione coerente di soluzioni di IA in modo scalabile negli ambienti di cloud ibrido e all'edge della rete.

Le operazioni IT possono beneficiare di configurazioni semplificate e flussi di lavoro più automatizzati su una piattaforma collaudata, scalabile con il minimo sforzo, che garantisce al contempo una governance e una sicurezza migliori.

Conquista la flessibilità del cloud ibrido

Red Hat OpenShift AI consente l'addestramento, il deployment e il monitoraggio dei carichi di lavoro AI/ML in vari ambienti (cloud, datacenter on premise o ambienti isolati) per soddisfare i requisiti normativi, di sicurezza e di dati. La piattaforma è compatibile con numerosi acceleratori di IA di fornitori come NVIDIA, AMD e Intel. È possibile espandere questa funzionalità per creare un ambiente GPU-as-a-service, che consente alle organizzazioni di gestire, partizionare e pianificare centralmente le risorse GPU, fornendo al contempo un'osservabilità dettagliata del loro utilizzo.

IA gen e Agentic AI

Per i progetti di IA gen, vengono offerte esperienze utente dedicate attraverso componenti come AI hub (Developer Preview), una dashboard per gli ingegneri di piattaforma che consolida catalogo, registro e deployment dei modelli per configurare e distribuire sia i modelli stessi che i server MCP. Gen AI studio (Developer Preview) fornisce endpoint per le risorse IA e un ambiente di test interattivo dove ingegneri di IA e sviluppatori di applicazioni possono accedere, sperimentare, confrontare e testare i modelli distribuiti e i server MCP.

OpenShift AI accelera Agentic AI fornendo un livello API unificato e una base flessibile e scalabile. Il supporto per Llama Stack API e MCP (Tech Preview) include un'implementazione di livello enterprise dell'API Llama Stack, la quale offre un unico punto di accesso standardizzato per varie funzionalità di IA.

Tra gli strumenti aggiuntivi figurano la valutazione degli LLM (LM Eval) e il benchmarking degli LLM a supporto dei deployment di inferenza nel mondo reale. LLM compressor fornisce algoritmi per ridurre le dimensioni dei modelli personalizzati di un'organizzazione utilizzando metodi simili a quelli impiegati da Red Hat per creare modelli convalidati e ottimizzati nel repository Red Hat AI su Hugging Face.

---

1 *IDC Tech Supplier. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Doc #US53418426, Oct. 2025. (richiede l'accesso al client)*

Red Hat
OpenShift AI

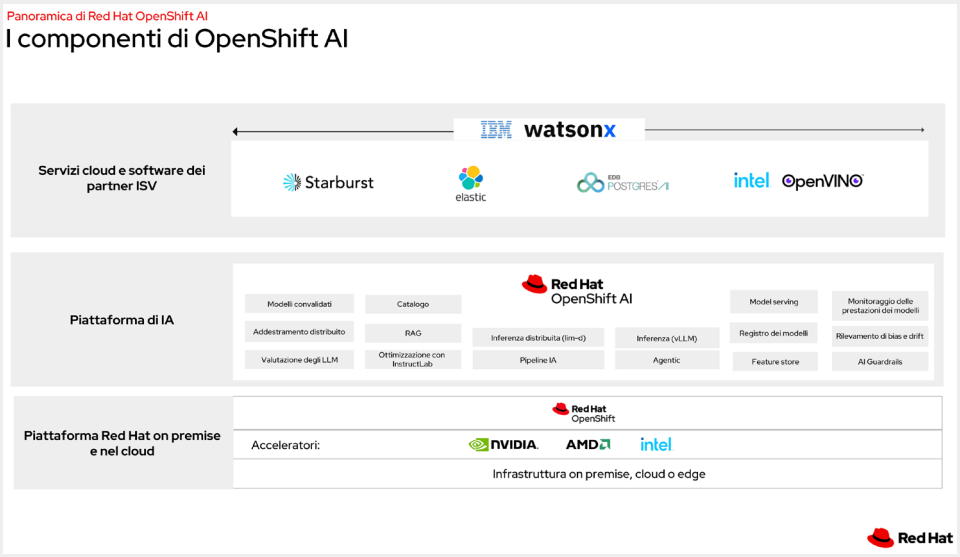

Figura 1. Red Hat OpenShift AI e componenti opzionali dei partner

Red Hat OpenShift AI offre una base solida grazie a diversi strumenti e funzionalità chiave:

- **Creazione e personalizzazione dei modelli.** I data scientist possono condurre analisi esplorative in un'interfaccia utente JupyterLab, che offre in dotazione immagini dei notebook sicure e integrate con le librerie Python più conosciute. Per i progetti di IA gen, OpenShift AI abilita la retrieval augmented generation (RAG) e l'addestramento distribuito InstructLab, fornendo strumenti per l'allineamento dei modelli per contribuire in modo più efficiente alle competenze e alle conoscenze dei modelli di IA gen.
- **Model serving.** Red Hat OpenShift AI mette a disposizione un'ampia gamma di framework che usano KServe come motore principale per il model serving, per semplificare il deployment di modelli predittivi di machine learning o modelli fondativi negli ambienti di produzione. Per gli LLM che necessitano di elevata scalabilità, OpenShift AI offre un servizio parallelo con runtime vLLM. Llm-d offre un framework per l'ottimizzazione dell'inferenza LLM disaggregando la pipeline in servizi modulari, il che supporta la scalabilità automatica intelligente e un routing efficiente delle richieste.
- **Pipeline AI.** Red Hat OpenShift AI offre un componente dedicato alle pipeline che consente di organizzare le attività di IA in pipeline e di creare pipeline utilizzando un'interfaccia grafica. Le organizzazioni possono concatenare i processi come la preparazione dei dati, la creazione dei modelli e la distribuzione dei modelli in produzione.
- **Monitoraggio dei modelli.** Red Hat OpenShift AI permette ai team operativi di monitorare le attività e le metriche prestazionali relative ai server dei modelli e ai modelli distribuiti. Gli utenti possono accedere a visualizzazioni pronte all'uso per le metriche relative alle prestazioni e alle operazioni oppure integrare i dati con altri servizi di osservabilità.
- **Carichi di lavoro distribuiti.** I carichi di lavoro distribuiti permettono di accelerare l'elaborazione dei dati, l'addestramento, il fine tuning e il model serving. Questa funzionalità consente anche di gerarchizzare e distribuire l'esecuzione dei processi e assicurare l'uso ottimale dei nodi. Il supporto avanzato per le GPU aiuta a gestire le esigenze dei carichi di lavoro relativi ai modelli fondativi.

- **Rilevamento di AI guardrail, bias, e drift.** Red Hat OpenShift AI offre strumenti volti ad aiutare i data scientist e gli ingegneri di IA a verificare il grado di equità e imparzialità dei modelli sulla base dei dati di addestramento, garantendo inoltre l'equità durante i deployment nel mondo reale. Gli AI guardrail forniscono un framework personalizzabile che implementa controlli di sicurezza fondamentali, contribuendo a garantire che i modelli siano trasparenti, equi e affidabili per l'uso in produzione. Gli strumenti per il rilevamento dei drift includono distribuzioni di dati di input per i modelli di ML di cui viene eseguito il deployment. In questo modo permettono di individuare il momento in cui i dati in tempo reale utilizzati per l'inferenza iniziano a discostarsi in modo significativo dai dati su cui è stato addestrato il modello.
- **Catalogo e registro.** Red Hat OpenShift AI mette a disposizione un catalogo di modelli interni e un catalogo di elementi selezionati in cui gli ingegneri della piattaforma possono individuare, confrontare e valutare modelli di IA gen ottimizzati. Inoltre, offre un registro centrale che consente ai data scientist e agli ingegneri di IA di condividere, modificare, eseguire il deployment e monitorare modelli predittivi e di IA gen, metadati e artefatti dei modelli.
- **Feature store.** Consente di gestire feature di dati pulite e ben definite per i modelli di ML, migliorando le prestazioni e accelerando i flussi di lavoro.

Strumenti per l'intero ciclo di vita dell'IA

Red Hat OpenShift fornisce tutti i servizi e i software necessari per l'addestramento e il deployment dei modelli e il loro trasferimento alla fase di produzione (vedi Figura 2).

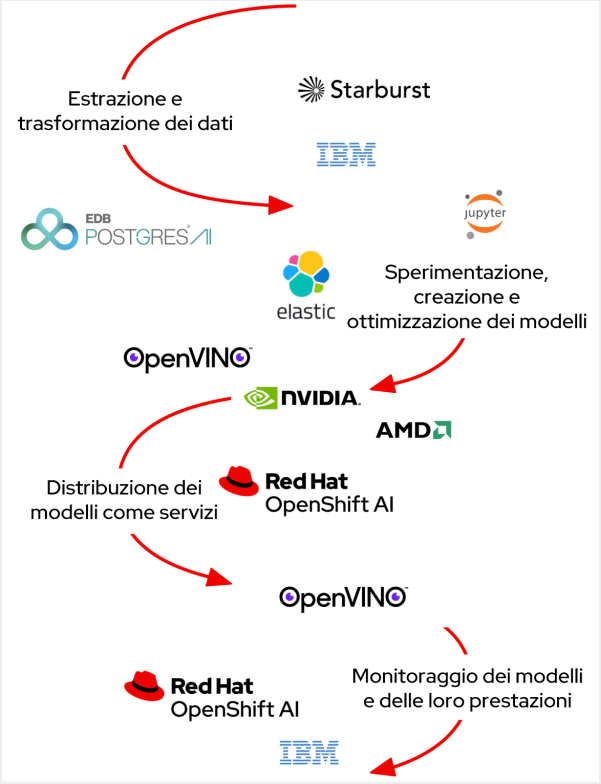

Figura 2. Il ciclo di vita standard per la realizzazione di un modello

La dashboard di Red Hat OpenShift AI facilita l'implementazione e consente di scoprire tutte le applicazioni e la documentazione e di accedervi da un'unica posizione. I tutorial Smart Start suggeriscono le procedure consigliate per i componenti e i principali software dei partner integrati nella piattaforma. È possibile accedervi direttamente dalla dashboard affinché i data scientist possano trovare agevolmente il supporto di cui hanno bisogno durante le fasi iniziali. Le sezioni seguenti descrivono gli strumenti dei partner tecnologici integrati in Red Hat OpenShift AI. Per utilizzare alcuni di questi strumenti occorre acquistare una licenza aggiuntiva tramite i canali del partner tecnologico.

Starburst

Starburst accelera le attività di analisi perché consente ai team di utilizzare i dati in modo rapido e veloce per migliorare le operazioni aziendali. Starburst viene fornito come prodotto completamente gestito o gestito in modo autonomo. È una soluzione che democratizza l'accesso ai dati, offrendo informazioni complete. Starburst è basato sul motore open source Trino (precedentemente noto come PrestoSQL), il primo motore SQL (Structured Query Language) per l'elaborazione massivamente parallela (MPP). Sviluppato e gestito dagli esperti di Trino, Starburst consente di interrogare diversi set di dati, ovunque essi si trovino, senza doverli spostare.

Starburst si integra con i servizi di elaborazione e storage su cloud a elevata scalabilità forniti da Red Hat OpenShift e consente di interrogare tutti i dati aziendali in modo affidabile, sicuro, efficiente e conveniente. I vantaggi includono:

- **Automazione.** Gli operatori Starburst e Red Hat OpenShift consentono di automatizzare la configurazione, l'ottimizzazione e la gestione dei cluster.
- **Alta disponibilità e scalabilità graduale.** Il bilanciamento del carico di Red Hat OpenShift consente di mantenere sempre disponibili servizi come il coordinatore Trino.
- **Scalabilità flessibile.** Red Hat OpenShift può scalare automaticamente i cluster di lavoro Trino in base al carico di query.

HPE Machine Learning Data Management Software

Le imprese hanno bisogno di soluzioni di gestione dei dati capaci di semplificare tutte le attività, dagli esperimenti individuali ai deployment aziendali di importanza critica. HPE Machine Learning Data Management Software (noto in precedenza come Pachyderm) consente ai team di scienza dei dati di creare pipeline di ML containerizzate, basate sui dati e scalabili, con un data lineage assicurato dal versioning automatico dei dati. Progettato per risolvere problemi reali nel campo della scienza dei dati, HPE Machine Learning Data Management Software fornisce la base dati necessaria ai team per automatizzare e scalare il ciclo di vita del machine learning con riproducibilità. Gli scenari di utilizzo includono dati non strutturati, data warehouse, elaborazione del linguaggio naturale, estrazione, trasformazione e caricamento (ETL) di video e immagini, servizi finanziari e bioscienze. HPE Machine Learning Data Management Software offre:

- Versioning automatico che consente ai team un monitoraggio ad alte prestazioni delle modifiche dei dati.
- Pipeline containerizzate e basate sui dati che accelerano l'elaborazione dei dati e riducono i costi.
- Data lineage immutabile che fornisce un registro fisso delle attività e delle risorse del ciclo di vita del machine learning.
- Una console che offre una visualizzazione intuitiva del grafo aciclico diretto (DAG) e un supporto per il debug e la riproducibilità.
- Supporto per i notebook Jupyter grazie a JupyterLab Mount Extension. L'estensione offre un'interfaccia intuitiva con cui accedere alle versioni dei dati.

- Strumenti affidabili per la gestione e il deployment scalabile di HPE Machine Learning Data Management Software nei diversi team aziendali.

NVIDIA accelera il deployment delle soluzioni di IA

Oggi che le applicazioni di AI/ML hanno un ruolo decisivo per il successo aziendale, le organizzazioni necessitano di piattaforme in grado di gestire carichi di lavoro complessi, ottimizzare l'uso delle risorse hardware e garantire la scalabilità. Le attività computazionali a elevato utilizzo di risorse includono l'inferenza, l'addestramento del ML, l'analisi dei dati e la loro elaborazione scalabile. Il software NVIDIA consente di accelerare tutti gli aspetti della scienza dei dati end to end sfruttando le funzionalità di elaborazione in parallelo delle GPU.

NVIDIA NIM migliora la gestione e le prestazioni delle GPU NVIDIA nell'ambiente Red Hat OpenShift, consentendo alle applicazioni di IA di sfruttare appieno il potenziale del software e dell'hardware di IA di NVIDIA. L'integrazione di NVIDIA NIM e Red Hat OpenShift AI consente una migliore allocazione delle risorse, una maggiore efficienza e un'esecuzione più produttiva dei carichi di lavoro dell'IA.

Toolkit Intel OpenVINO

Il toolkit Intel OpenVINO accelera lo sviluppo e il deployment delle applicazioni di inferenza per il DL ad alte prestazioni sulle piattaforme Intel. Il toolkit consente di adottare, ottimizzare ed eseguire il fine tuning sui modelli di rete neurale virtualmente e svolgere attività di inferenza complete utilizzando l'ecosistema di strumenti di sviluppo OpenVINO.

- **Modello.** Gli sviluppatori di software possono utilizzare i propri modelli di DL. Per accelerare i tempi di rilascio, possono anche servirsi dei modelli preaddestrati e preottimizzati presenti nel toolkit OpenVINO grazie alla collaborazione tra Intel e Hugging Face. OpenVINO supporta Pytorch, ONNX, TensorFlow e altri formati di modelli diffusi.
- **Ottimizzazione.** Il toolkit OpenVINO offre diverse modalità per convertire i modelli e aiutare gli sviluppatori di software a ottenere modelli di IA più veloci ed efficienti. Gli sviluppatori possono comunque saltare la conversione dei modelli ed eseguire l'inferenza direttamente nei formati PyTorch, ONNX, TensorFlow, TensorFlow Lite, JAX, o PaddlePaddle. La conversione in OpenVINO IR garantisce prestazioni ottimali, che possono essere ulteriormente migliorate utilizzando le funzionalità di compressione dei pesi e quantizzazione disponibili nel Neural Network Compression Framework di OpenVINO. Le stesse funzionalità riducono anche l'impatto dello storage e del runtime.
- **Deployment.** OpenVINO Runtime Inference Engine è un'API progettata per essere integrata nelle applicazioni e accelerare i processi di inferenza. Il suo approccio "scrivi un volta, esegui il deployment ovunque" consente di eseguire in modo efficiente attività di inferenza sui diversi hardware Intel, tra cui CPU, GPU, NPU e FPGA. La libreria di estensioni GenAI di OpenVINO semplifica il deployment dei carichi di lavoro dell'IA gen, in molti casi riducendo il codice necessario a sole 3-5 righe. OpenVINO Model Server offre numerose funzionalità per gli scenari Agentic e model serving, riducendo ulteriormente le attività di sviluppo.

EDB

EDB Postgres AI è una piattaforma potente e intelligente progettata per gestire i carichi di lavoro transazionali, analitici e di IA, che offre una flessibilità senza precedenti a prescindere dal fatto che i dati risiedano on premise o in un qualsiasi ambiente cloud. Leader mondiale nelle soluzioni di database Postgres per aziende, EDB offre una piattaforma aperta, di livello enterprise, per la gestione dei dati e l'IA che contribuisce ad accelerare fino a tre volte la produzione dei progetti di IA. Grazie all'integrazione con Red Hat OpenShift AI, EDB Postgres AI consente agli utenti di creare solide basi di conoscenza dell'IA per la retrieval augmented generation (RAG), unificando dati, modelli e applicazioni dell'IA in una piattaforma di IA sovrana completa e distribuibile

ovunque. La trasformazione dei dati operativi di base in una risorsa predisposta per l'IA può aumentare l'efficienza fino al 30%, oltre a semplificare l'uso dei dati privati, inclusi i dati non strutturati, per integrare gli output dei modelli nella knowledge base di un'organizzazione.

Elastic

Elastic Search AI Platform (sviluppata a partire da ELK Stack[2]) combina la precisione della ricerca e l'intelligenza dell'IA per consentire agli utenti di creare prototipi e integrarli negli LLM più velocemente e di utilizzare l'IA gen per la creazione a costi ridotti di applicazioni scalabili. Elastic Search AI Platform permette di creare applicazioni di retrieval augmented generation (RAG), di risolvere in modo proattivo i problemi di osservabilità e di affrontare anche le più complesse minacce alla sicurezza. Inoltre, si può distribuire in qualunque ambiente: on premise, nel cloud e in ambienti isolati.

Elastic si integra con i modelli di embedding dell'ecosistema, che comprende Red Hat OpenShift AI, Hugging Face, Cohere, OpenAI, tramite un'unica chiamata API intuitiva. Questo approccio permette di ottenere un codice adatto alla gestione dell'inferenza ibrida dei carichi di lavoro RAG, con funzionalità che includono:

- Suddivisione, connettori e web crawler per l'inserimento di diversi set di dati nel livello di ricerca.
- Ricerca semantica grazie a Elastic Learned Sparse EncodeR (ELSER), il modello di ML integrato, e il modello di embedding E5 al fine di permettere la ricerca vettoriale multilingue.
- Sicurezza a livello di documento e ambito grazie ad autorizzazioni in linea con i criteri di controllo degli accessi basato sui ruoli (RBAC) dell'organizzazione.

Chi adotta Elastic Search AI Platform entra a far parte di una community mondiale di sviluppatori esperti con offrono ispirazione e supporto. Scopri la community Elastic su Slack, i forum di discussione e gli account social.

Conclusioni

Con Red Hat OpenShift AI, le imprese possono sperimentare, collaborare e accelerare il percorso verso lo sviluppo di applicazioni basate sull'IA. I data scientist e gli ingegneri di IA acquisiscono la flessibilità necessaria per utilizzare Red Hat OpenShift AI per creare e implementare modelli nel cloud ibrido. Gli ingegneri delle operazioni IT e delle piattaforme si avvalgono delle funzionalità MLOps e GenAIOps, che consentono di eseguire il deployment dei modelli in produzione più rapidamente. L'accesso self service per sviluppatori, ingegneri di IA e data scientist, compreso l'accesso alle GPU, stimola l'innovazione su una piattaforma applicativa già utilizzata e pienamente affidabile per l'IT aziendale. Red Hat OpenShift AI continua a offrire una piattaforma completa, affidabile e coerente, che si distingue con elementi unici in termini di inferenza efficiente, Agentic AI e operazioni di cloud ibrido scalabili, supportate da un solido ecosistema di partner.

Scopri di più

Inizia oggi stesso visitando Red Hat OpenShift AI.

Informazioni su Red Hat

Red Hat consente la standardizzazione in diversi ambienti e lo sviluppo di applicazioni cloud native, oltre a favorire l'integrazione, l'automazione, la protezione e la gestione di ambienti complessi grazie a pluripremiati servizi di consulenza, formazione e supporto.

2 *Lo stack ELK è composto da Elasticsearch, Kibana, Beats e Logstash.*

facebook.com/RedHatItaly
twitter.com/RedHatItaly
linkedin.com/company/red-hat

ITALIA
it.redhat.com
italy@redhat.com

EUROPA, MEDIO ORIENTE, E AFRICA (EMEA)
00800 7334 2835
it.redhat.com
europe@redhat.com

it.redhat.com
#2876428_1125