Red Hat accelera l'adozione dell'IA per i servizi finanziari

Riduci i tempi di rilascio per le soluzioni di AI/ML con una piattaforma end to end

La maggiore complessità dei modelli di IA rende più difficoltosa l'adozione

Gli istituti finanziari puntano ad adottare l'intelligenza artificiale (IA) in cerca di opportunità di monetizzazione. La rapida evoluzione di deep learning, IA conversazionale e IA generativa ha sensibilmente ampliato l'ambito di applicazione delle soluzioni di IA. Al contempo però la crescente complessità dei modelli genera nuove problematiche legate all'esecuzione, acuendo le difficoltà già esistenti. Di seguito alcune delle sfide principali:

- **Processo di sviluppo standalone:** la maggior parte delle attività di sviluppo e addestramento dell'intelligenza artificiale e del machine learning (ML) si svolge oggi in ambienti dedicati e necessita di risorse speciali, ad esempio acceleratori hardware come le unità di elaborazione grafica (GPU). Il provisioning degli ambienti di AI/ML richiede tempo e rallenta il rilascio di nuovi servizi basati sull'IA.
- **Scalabilità, flessibilità e ottimizzazione delle risorse:** le soluzioni di AI/ML necessitano di componenti che hanno esigenze diverse in termini di risorse, ad esempio l'unità di elaborazione centrale (CPU), la memoria, il disco e l'hardware specializzato (GPU), l'unità di elaborazione del tensore (TPU) e il circuito FPGA (Field-Programmable Gate Array). Per assicurare la scalabilità di queste soluzioni occorre un approccio di cloud ibrido.
- **Monitoraggio e deviazioni:** i modelli di AI/ML devono essere monitorati in continuo e aggiornati regolarmente per rilevare e correggere le deviazioni. Red Hat® OpenShift® agevola l'integrazione continua degli aggiornamenti dei modelli perché offre un'infrastruttura di monitoraggio basata standard che collega il monitoraggio delle deviazioni incentrato sulle applicazioni alla pipeline di sviluppo dell'AI/ML.
- **Sicurezza della catena di distribuzione dei modelli:** l'ecosistema di strumenti per lo sviluppo dell'AI/ML offre per la maggior parte framework open source sviluppati dalla community. Garantire l'integrità della catena di distribuzione del software in questo contesto sta diventando sempre più difficile. Se da un lato gli sviluppatori vorrebbero utilizzare sempre gli strumenti più recenti, dall'altro le organizzazioni devono garantire che tali strumenti siano sicuri e non contengano artefatti dannosi o vulnerabili.

Vantaggi che riducono in maniera significativa le complessità

La soluzione di AI/ML proposta per gli istituti finanziari offre i seguenti vantaggi:

- Una piattaforma end to end per lo sviluppo, l'addestramento e l'inferenza dei modelli. Migliora la coerenza delle operazioni tra gli ambienti cloud pubblici e privati, allineando al contempo le diverse fasi del processo.
- Funzionalità self service che accelerano il time to value degli ambienti di ML.
- Un insieme coerente e aggiornato di strumenti e librerie di ML open source, oltre a un vasto ecosistema di tecnologie open source e supportate da partner certificati.
- Sviluppo e deployment rapidi dei modelli di ML e funzionalità di monitoraggio e iterazione rapida per assicurare l'aggiornamento costante dei modelli.

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

Modelli linguistici di grandi dimensioni

Un esempio che illustra i vantaggi e le sfide nel settore dei servizi finanziari è l'implementazione di soluzioni basate su modelli linguistici di grandi dimensioni (LLM), come GPT-4, BLOOM, BART, DOLLY e altri. Queste soluzioni vengono utilizzate per digitalizzare i documenti durante i processi di onboarding o di conoscenza del cliente, analizzare i report sui dati ambientali, sociali e di governance aziendale (ESG) o implementare soluzioni conversazionali, come i chatbot.

Queste soluzioni si basano in genere su modelli di ML di grandi dimensioni con centinaia di milioni o miliardi di parametri. A causa delle varie fasi coinvolte, della complessità e della potenza di elaborazione necessaria per creare questi modelli, è prassi comune servirsi di modelli preaddestrati o modelli fondativi. Si tratta di modelli addestrati su set di dati per utilizzo generico, questo significa che per applicarli agli scenari di utilizzo dei servizi finanziari occorre eseguire un addestramento aggiuntivo specifico per dominio o azienda su un insieme più piccolo di dati locali, con tecniche quali fine tuning o transfer learning. La Figura 1 illustra un esempio di architettura per questo tipo di soluzione.

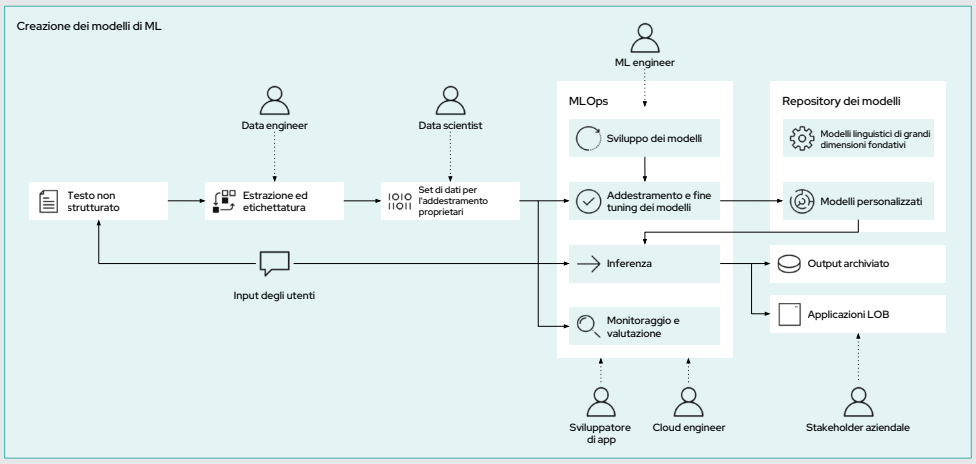

Figura 1. Panoramica sulla creazione dei modelli di ML

Panoramica sulle funzionalità

Architettura della soluzione

Red Hat offre una piattaforma capace di ospitare in maniera efficiente ed efficace l'intero ciclo di vita dell'AI/ML, dallo sviluppo all'addestramento all'inferenza. La piattaforma Red Hat si può eseguire su tutte le principali infrastrutture, bare metal, virtualizzazione on premise e i principali cloud pubblici. Questo significa che le organizzazioni possono utilizzare la stessa piattaforma, gli stessi strumenti e gli stessi processi MLOps in tutti gli ambienti.

Consapevole delle potenzialità dell'open source ma anche dell'importanza di garantire una catena di distribuzione sicura, Red Hat partecipa attivamente alle community upstream dove contribuisce allo sviluppo di software innovativi, instaurando strette collaborazioni. In questo senso, Red Hat è impegnata a selezionare, supportare e certificare tutta una serie di strumenti upstream necessari agli sviluppatori dell'AI/ML. Lascia che sia Red Hat ad analizzare la catena di distribuzione upstream e a fornire alla tua azienda un prodotto di livello enterprise affidabile e capace di garantire un supporto sempre disponibile.

Componenti della piattaforma

Sistema operativo

L'architettura di AI/ML di Red Hat si basa su Red Hat Enterprise Linux®, un sistema operativo che si può eseguire su infrastrutture on premise con deployment moderni, in ambienti cloud, bare metal e su macchine virtuali. Red Hat Enterprise Linux è certificato per l'utilizzo con un ampio ecosistema di hardware e con i principali provider di servizi cloud, tra cui Amazon Web Service (AWS), Google Cloud, IBM Cloud for Financial Services, Oracle Cloud e Microsoft Azure. La piattaforma Linux garantisce sicurezza, prestazioni elevate e supporto, oltre a funzionalità di automazione avanzate tramite Red Hat Ansible® Automation Platform. Red Hat Enterprise Linux offre anche il supporto per gli hardware specializzati necessari allo sviluppo dell'AI/ML, come GPU e FPGA.

Orchestrazione dei container

Oltre alle applicazioni personalizzate e quelle disponibili in commercio, la maggior parte degli strumenti e delle librerie open source utilizzati nei processi di AI/ML è containerizzata. Anche i modelli di ML preaddestrati o di produzione vengono forniti come immagini dei container. Inoltre, per ottenere processi di AI/ML efficaci è necessario assicurare l'interazione corretta e la scalabilità dei diversi componenti coinvolti. Adottare una piattaforma flessibile e scalabile è quindi indispensabile per la gestione di componenti quali l'addestramento intensivo dei nuovi modelli, i motori inferenziali ad elevata velocità di elaborazione e gli ambienti di sviluppo dei modelli utilizzati dai data scientist. La piattaforma leader di settore per il deployment e l'orchestrazione dei carichi di lavoro containerizzati è Red Hat OpenShift, una distribuzione Kubernetes. È la piattaforma più utilizzata per gli strumenti di sviluppo dell'IA di terze parti e open source e garantisce ai team di sviluppo l'accesso ai framework di AI/ML di cui hanno bisogno per accelerare il time to value. Red Hat OpenShift offre anche degli operatori per automatizzare il deployment dei componenti. Un'opzione utile che incentiva l'approccio self service e riduce i costi operativi.

Storage scalabile e sicuro

Nei progetti di AI/ML servono grandi quantità di dati per creare modelli accurati. I dati possono essere storici o disponibili live da sorgenti quali flussi di dati di mercato, Internet of Things (IoT) e osservabilità. Quale che sia la loro origine, lo storage deve essere intuitivo e di facile accesso per gli sviluppatori. Red Hat supporta e integra Red Hat OpenShift Data Foundation, una soluzione di storage software defined open source basata su Red Hat Ceph® Storage. OpenShift Data Foundation si integra con Red Hat OpenShift e assicura la scalabilità nell'ordine dei petabyte e oltre ad un costo contenuto. Inoltre, AMQ Streams, una soluzione basata su Apache Kafka, garantisce agli sviluppatori l'accesso ripetuto ai dati in streaming. Entrambe le soluzioni, OpenShift Data Foundation e AMQ Streams, sono predisposte in container e gestibili con Red Hat OpenShift. In questo modo più team di sviluppo potranno utilizzarle in modalità self service.

Funzionalità della piattaforma

Self service

Red Hat OpenShift permette di eseguire l'onboarding on demand dei team di sviluppo e dei progetti, assicura la scalabilità delle risorse e consente di raccogliere in pool e condividere i costosi componenti hardware specializzati, come le GPU. Inoltre, integra la conformità e la sicurezza della catena di distribuzione del software a tutti i livelli.

Osservabilità e monitoraggio avanzati

Red Hat OpenShift offre funzionalità di monitoraggio open source di alto livello grazie a Prometheus ed è compatibile con numerosi strumenti di monitoraggio di terze parti, come Splunk. Permette di integrare le pipeline MLOps in un'infrastruttura centralizzata flessibile che assicura monitoraggio e notifica tempestiva dei problemi in tutta la pipeline. Tenere traccia delle prestazioni dei modelli consente di automatizzare la scalabilità e la generazione di avvisi quando si riscontrano livelli di accuratezza bassi.

Agilità

La creazione dei modelli di AI/ML è un processo iterativo. I professionisti del settore, come data engineer e data scientist, analizzano i percorsi lasciati dalle tracce dei dati. Il percorso che porta allo sviluppo del modello è spesso soggetto a interruzioni e riprese, con complicazioni impreviste e interventi da cui recedere. Gli ostacoli più comuni includono l'accesso a dati di qualità provenienti da sorgenti diverse, come database, file system, flussi, interfacce di programmazione delle applicazioni (API), e la conformità agli obblighi normativi e agli standard di sicurezza. Per quanto riguarda gli strumenti, le sfide includono il controllo delle versioni su un'ampia gamma di librerie, l'aggiornamento degli strumenti esistenti e l'adozione di nuove soluzione. Red Hat aiuta a semplificare la pipeline di AI/ML offrendo ai professionisti un'esperienza coerente nell'intero ambiente cloud ibrido, e contribuisce così ad accelerare i progetti di AI/ML.

Una differenza tra lo sviluppo di applicazioni tradizionali e lo sviluppo di applicazioni di AI/ML risiede nella necessità assoluta di aggiornare le applicazioni e i modelli di IA alla base di tali applicazioni. Oltre allo sviluppo iniziale dei modelli tramite ML, le tecniche di AI/ML devono permettere anche l'aggiornamento continuo dei modelli. In questo modo i modelli sono in grado di offrire vantaggi nettamente superiori alle applicazioni tradizionali, ma occorrono interventi e aggiornamenti costanti sui modelli per migliorare le prestazioni. Red Hat OpenShift assicura ai team dedicati alle applicazioni la possibilità di ampliare o diminuire la portata dei componenti della toolchain MLOps. Quando un'applicazione richiede l'aggiornamento del modello, è possibile assegnare ed espandere manualmente le risorse per l'addestramento (più costose), come GPU e altri componenti specializzati. Una volta ultimato l'aggiornamento, Red Hat OpenShift riassegna tali risorse dove servono.

Scalabilità ed elasticità per l'addestramento e l'inferenza

L'addestramento all'interno del processo di creazione dei modelli di AI/ML è una delle operazioni che richiedono più risorse nella pipeline MLOps. È in questa fase che più facilmente si estende la portata degli strumenti di AI/ML e che la domanda di hardware specializzati, come GPU, TPU e FPGA, da parte di aziende come Nvidia è più elevata. Per svolgere l'addestramento, i singoli progetti e team devono accedere ai propri ambienti. La capacità dell'architettura di AI/ML di Red Hat di offrire un'infrastruttura condivisa garantisce notevoli vantaggi in termini di efficienza e riduzione dei costi. Anziché accumulare costose risorse dedicate, con Red Hat OpenShift gli sviluppatori possono accedere in modo virtuale e on demand all'intero cluster. Kubernetes si occupa di orchestrare tali accessi e assicura che le risorse vengano distribuite dove occorre.

Ecosistema aperto

Come tutti gli altri prodotti Red Hat, anche la piattaforma di AI/ML è una soluzione open source. L'ecosistema open source di strumenti e tecnologie disponibile per i professionisti dell'AI/ML include:

- Librerie di ML.
- Gestione del ciclo di vita dell'AI/ML.
- Accesso ai dati, qualità dei dati e gestione dei metadati.
- Rilevamento dei pregiudizi ed esplicabilità.
- Modelli preaddestrati.

Grazie all'accesibilità dell'ecosistema open source e alla flessibilità della piattaforma, questi strumenti si possono utilizzare insieme in varie combinazioni, a seconda delle soluzioni. Inoltre, disporre di una piattaforma aperta agevola l'innovazione continua e consente l'integrazione delle tecnologie, degli strumenti e dei modelli più recenti.

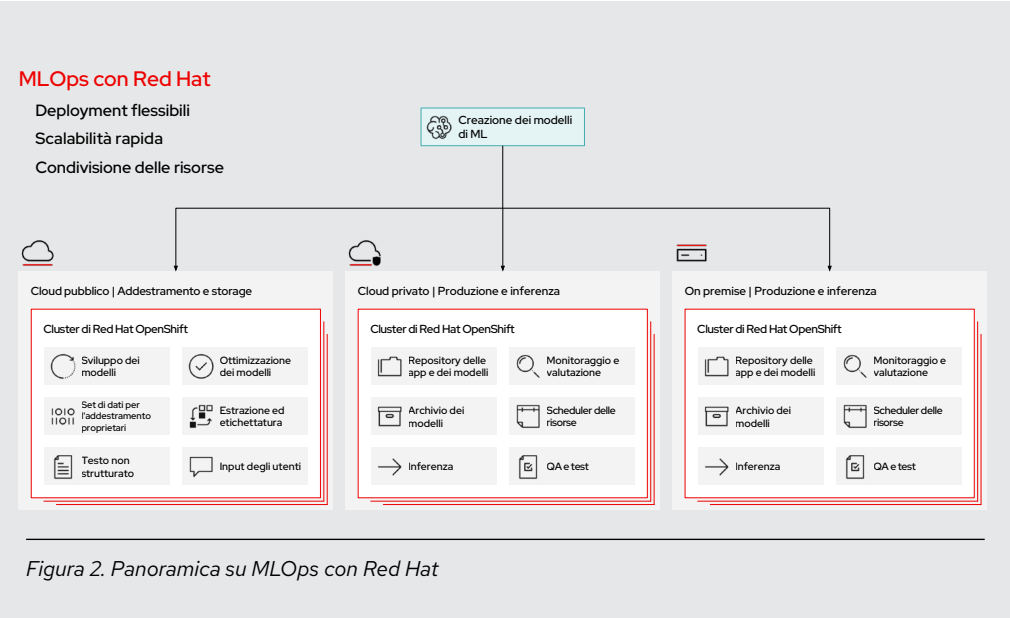

Figura 2. Panoramica su MLOps con Red Hat

Informazioni su Red Hat

Red Hat è leader mondiale nella fornitura di soluzioni software open source di livello enterprise. Con un approccio basato sul concetto di community, distribuisce tecnologie come Kubernetes, container, Linux e cloud ibrido caratterizzate da affidabilità e prestazioni elevate. Red Hat consente di sviluppare applicazioni cloud native, integrare applicazioni IT nuove ed esistenti, e automatizzare e gestire ambienti complessi. Considerata un partner affidabile dalle aziende della classifica Fortune 500, Red Hat fornisce pluripremiati servizi di consulenza, formazione e assistenza, che portano i vantaggi dell'innovazione open source in qualsiasi settore. Red Hat è l'elemento catalizzatore in una rete globale di aziende, partner e community, e permette alle organizzazioni di evolversi e prepararsi a un futuro digitale.

ITALIA
it.redhat.com
italy@redhat.com

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

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

it.redhat.com
#420450_0723