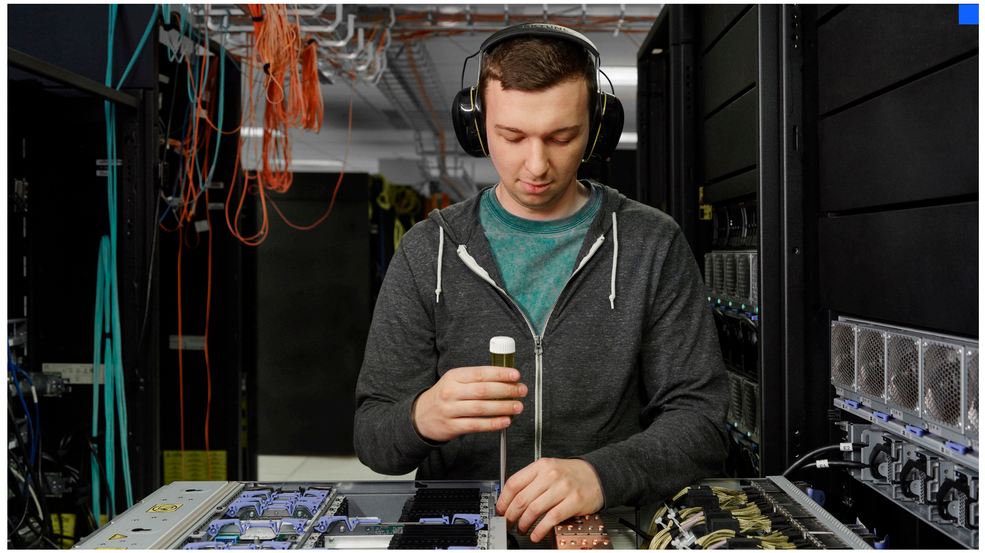
watsonx.data™

L' archivio dati

↳ per l'AI

IBM

Indice



01

Introduzione

Questo ebook esaminerà la più recente soluzione di gestione dei dati open source destinata a leader nel campo dei dati e dell'analytics, che vogliono ridurre in modo significativo i costi, semplificare l'accesso ai dati e automatizzare la governance unificata, per ridimensionare l'AI. È il momento di prendere in considerazione il data lakehouse.

I dati sono al centro di ogni attività aziendale. Mantengono in funzione le applicazioni, potenziano gli insight predittivi e consentono di offrire esperienze migliori a clienti e dipendenti. Ma, la comprensione totale del vantaggio che i dati possono offrire è difficilmente raggiungibile, a causa della modalità di archiviazione e accesso ai dati per operazioni di analytics e AI.

Non sei il solo ad affidarti a repository monolitici con numerosi data warehouse e data lake, on premise e su cloud; l'82% delle organizzazioni è ostacolato dall'utilizzo di silos di dati.[1] E la situazione è destinata a peggiorare: secondo IDC, si prevede che la quantità di dati archiviati crescerà del 250% entro il 2025.[2]

Il data lake avrebbe dovuto rappresentare la soluzione per tutti questi problemi; si posizionano semplicemente i dati in un'ubicazione centralizzata e si elaborano. Ma non è così facile aggiornare i lake, catalogare correttamente i dati o garantire una governance efficace—e le competenze richieste per queste attività sono specifiche, rare e costose. Ne consegue che la creazione e la manutenzione dei data lake si sono dimostrate costose. Un data warehouse offre prestazioni elevate per l'elaborazione di terabyte di dati strutturati. Ma anche i warehouse possono diventare costosi, soprattutto per carichi di lavoro nuovi e in evoluzione. La maggior parte delle organizzazioni esegue carichi di lavoro di analytics e AI in ecosistemi complessi e non redditizi. È il momento di cambiare.

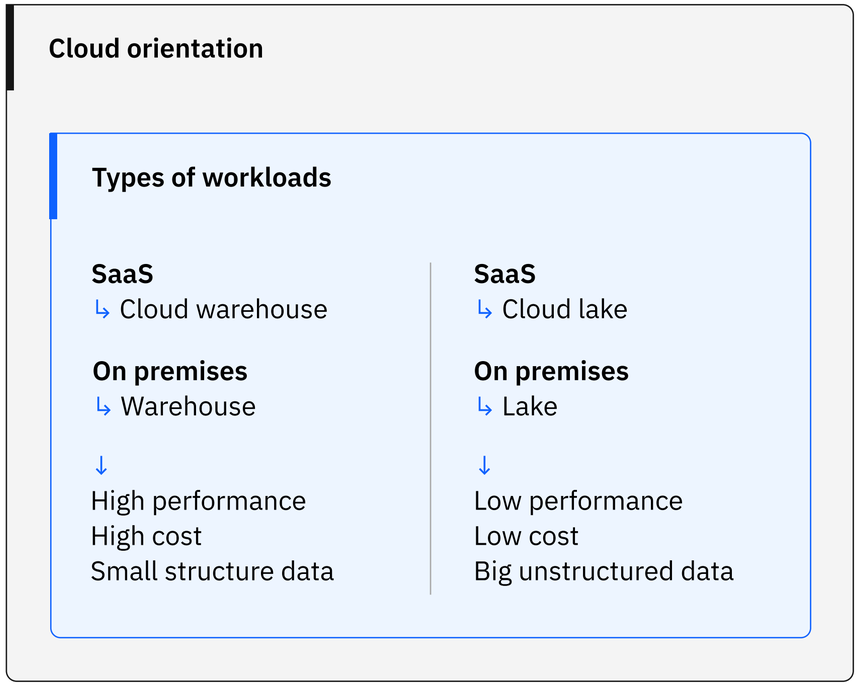
↑250%

La quantità di dati archiviati si prevede che crescerà del 250% entro il 2025.[2]

02

Lo stato corrente dell'architettura dei dati

Una combinazione di warehouse on premise e cloud-native e di data lake personalizzati oggi è una scelta comune nell'architettura aziendale. È probabile che la gestione dei costi, dei dati isolati in silos e della governance dei dati rappresenti una sfida costante.



Il data lakehouse è un cambiamento di paradigma emergente nel modo in cui le aziende ricavano insight.[3]

03

Il data lakehouse definito

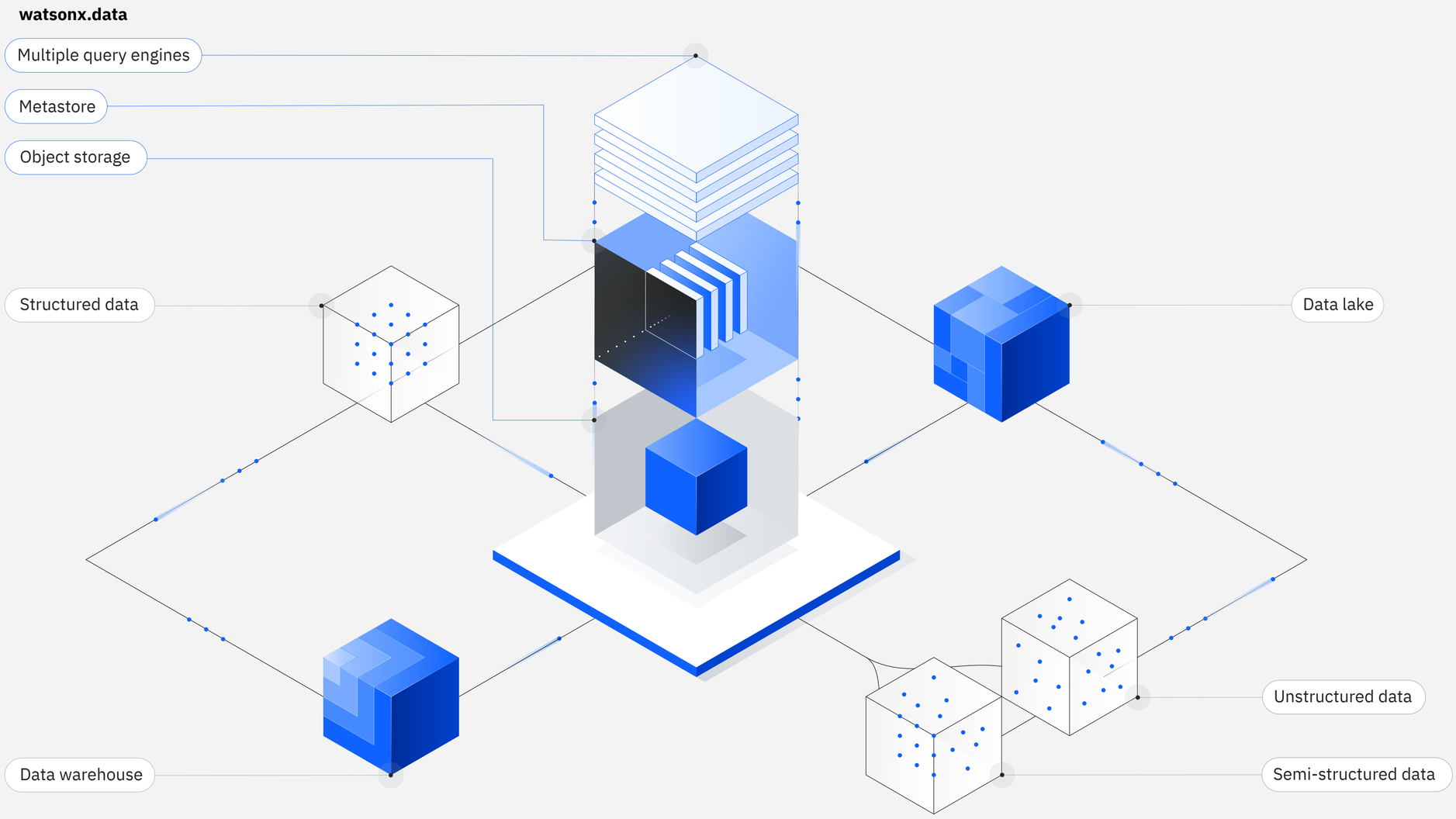
Cerca una soluzione lakehouse che fornisca una moderna base di dati per ridimensionare l'AI.

Il data lakehouse è un'architettura emergente che offre la flessibilità di un data lake con le prestazioni e la struttura di un data warehouse. La maggior parte delle soluzioni lakehouse offre un motore di query ad alte prestazioni, su uno storage a basso costo, insieme a un livello di governance dei metadati. Livelli di metadati intelligenti facilitano per gli utenti la categorizzazione e la classificazione dei dati non strutturati, ad esempio video e voce, e dei dati semi-strutturati, quali XML, JSON ed e-mail.

I migliori data lakehouse offriranno tecnologie open-source che riducono la duplicazione dei dati e semplificano le pipeline ETL complesse. Tieni presente che alcuni lakehouse di prima generazione presentano vincoli fondamentali che limitano la loro capacità di affrontare le sfide relative ai costi e alla complessità. Ad esempio, un singolo motore di query progettato per carichi di lavoro di business intelligence o ML (machine learning), potrebbe risultare inefficace, se utilizzato per un altro tipo di carico di lavoro.

Il team IBM per i dati e l'AI ritiene che ogni carico di lavoro sia unico e debba essere ottimizzato utilizzando l'ambiente più adatto, in grado di mantenere i costi al minimo e le prestazioni al massimo. Scegli un lakehouse che offra un livello ottimale di prestazioni per un processo decisionale più efficace, insieme alla flessibilità necessaria per ricavare il massimo valore da tutti i tipi di dati.

Figura 1. Come ridimensionare e accelerare al meglio l'impatto dell'AI

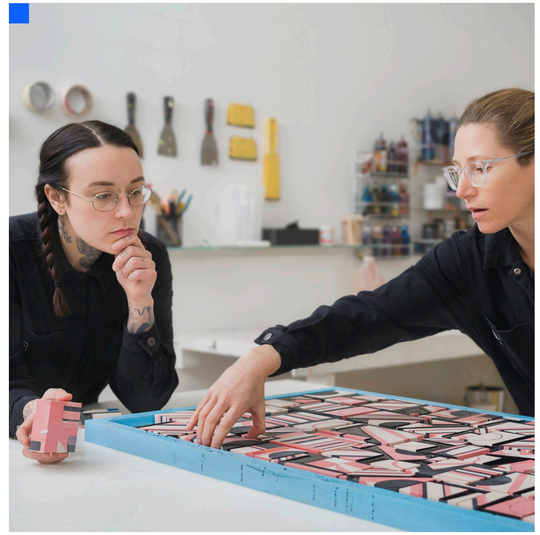


Componenti dell'architettura

Infrastruttura
Questo componente rappresenta l'ubicazione dove verrà implementato il lakehouse, completamente gestito su qualsiasi ambiente cloud oppure on-premise.

Storage
Questo livello rappresenta l'ubicazione dove vengono archiviati fisicamente i dati, che sono memorizzati come file e possono essere memorizzati in formati di dati aperti, come ad esempio Apache Parquet e Avro. I formati di dati aperti sono specifiche di file e protocolli resi disponibili alla community open-source, in modo che chiunque possa acquisirli e migliorarli.

Formati tabella aperti
I formati tabella aperti, quali Apache Iceberg, aiutano a fornire una struttura e offrono l'affidabilità e la semplicità di SQL con i big data. Questi formati consentono a motori diversi di accedere agli stessi dati, nello stesso momento, evitando così il vincolo imposto dal vendor. Condividi i dati tra più strumenti e repository di dati, ad esempio il tuo data warehouse; una singola copia dei dati ti permette di ridurre la duplicazione dei dati e di abbattere i silos.



Governance

Anche i metadati sono memorizzati con formati tabella aperti; servono a definire i formati di file per qualsiasi strumento in grado di leggere o scrivere formati di dati aperti.

Servizio di metadati tecnici
Questo componente è necessario per comprendere quali dati siano disponibili nel livello di storage. Il motore di query richiede i metadati per i dati e le tabelle per fornire una derivazione completa e per sapere dove sono ubicati, qual è il loro aspetto e come si leggono.

Cataloghi di dati
Questo componente aiuta gli utenti a trovare i dati corretti per il lavoro e fornisce informazioni semantiche per le politiche e le regole. Prevedi l'archiviazione di metadati di business, come ad esempio terminologie e tag di business per consentire la ricerca e la protezione dei dati.

Motore delle politiche
Questo componente consente agli utenti di definire le politiche di protezione dei dati e permette al motore di applicare tali politiche. Per creare un framework di governance che sia ridimensionabile, un motore delle politiche viene spesso implementato con il servizio di metadati tecnici e il catalogo dei dati.

Motore di query
Questo componente è al centro del lakehouse di dati aperto. Un motore di query, che può essere open source o proprietario, accede ai dati in formato tabella aperto ed è spesso noto come componente di calcolo. I motori di query sono generalmente di due tipi: un motore di query basato su SQL, ad esempio Presto open-source, o un motore Apache Spark open-source o un suo equivalente.

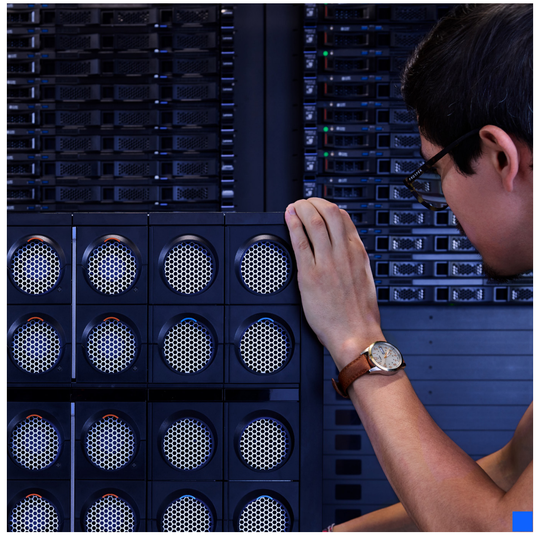
In un'architettura di lakehouse aperto, il motore di query è completamente modulare, il che significa che il motore può essere ridimensionato dinamicamente per fare fronte alle esigenze dei carichi di lavoro e alla concorrenza. I motori di query possono anche collegarsi a qualsiasi catalogo e storage.

↓50%

Ora è possibile ottenere insight più rapidi e attendibili, dimezzando i costi del data warehouse.[4]

05

Opportunità di ottimizzazione dei costi



Se la tua organizzazione dispone già di implementazioni di big data on premise, un lakehouse offre un'alternativa meno costosa per l'archiviazione dei dati in formati aperti in uno storage di oggetti. Ridurrai il costo dell'analytics, diminuirai la complessità e migliorerai il time-to-value.

Se esiste già un'implementazione warehouse, un approccio che prevede un lakehouse può rappresentare un'alternativa estremamente scalabile e a basso costo per carichi di lavoro di analytics di grandi dimensioni, meno sensibili agli SLA (service-level agreements). I warehouse sono spesso costosi e proprietari, ma, con un lakehouse, potrai drasticamente ridurre i costi di storage e di calcolo. Si possono ottimizzare i carichi di lavoro del warehouse utilizzando motori adatti allo scopo, che si basano sui requisiti del carico di lavoro. La natura aperta di un lakehouse ti libera dalla tecnologia warehouse proprietaria, il che significa minori vincoli imposti dal fornitore e una riduzione dei costi generali dell'infrastruttura IT.

IBM watsonx.data è un archivio dati aperto, ibrido e regolamentato, ottimizzato per tutti i carichi di lavoro di dati, analytics e AI.

06

Miglioramenti dell'analytics e della data science

"Ci stiamo muovendo nella direzione in cui il data lakehouse diventerà una best practice."[3]

Adam Ronthal
Vicepresidente
di Gartner

I formati di dati proprietari e gli elevati costi di storage limitano la collaborazione e le implementazioni di modelli AI e ML all'interno di un ambiente di data warehouse; i data lake vengono messi alla prova da carichi di lavoro di data science a basse prestazioni. L'isolamento di queste tecnologie ha portato a successive sfide per l'infrastruttura, unitamente alle implicazioni per sicurezza e governance che derivano dalla duplicazione e dallo spostamento dei dati per lo sviluppo di modelli di AI e ML.

Un data lakehouse è un ottimo modo per aiutare i colleghi ansiosi di ricevere gli insight che attendono di emergere dai dati della tua organizzazione. Se vuoi davvero ricavare valore di business dal flusso di dati in arrivo, prendi in considerazione l'adozione di una strategia di lakehouse.

Adam Ronthal, vicepresidente e analista presso Gartner, afferma “Ci stiamo muovendo nella direzione in cui il data lakehouse diventerà una best practice.”[2] L'approccio migliore offrirà un ambiente aperto, collaborativo e regolamentato per la gestione end-to-end dei carichi di lavoro di data science.

Esaminiamo IBM watsonx.data™—l'archivio dati aperto, ibrido e regolamentato, che è ottimizzato per tutti i carichi di lavoro di dati, analytics e AI.

07

IBM watsonx.data

Ridimensiona i carichi di lavoro AI, per tutti i tuoi dati, ovunque. Watsonx.data è un archivio dati aperto, ibrido e regolamentato, ottimizzato per tutti i carichi di lavoro di dati, analitycs e AI, basato su un'architettura di data lakehouse (vedere figura 1).

Accedi a tutti i tuoi dati e garantisci la massima copertura dei carichi di lavoro in tutti gli ambienti di cloud ibrido. Aspettati l'erogazione ininterrotta di un servizio completamente gestito in qualsiasi ambiente cloud oppure on-premise. Accedi a qualsiasi origine dati, ovunque risieda, tramite un singolo punto di ingresso e combina i dati usando formati di dati aperti. Esegui l'integrazione nel tuo ambiente esistente con open source e open standard e interoperabilità con servizi IBM e di terze parti.

Accelera i tempi per ottenere insight attendibili. Inizia subito con la governance e l'automazione integrate; rafforza la conformità e la sicurezza aziendale con una governance unificata in tutto l'ecosistema. Un'esperienza utente chiara e una console click-and-go aiutano i tuoi team ad acquisire, visualizzare e trasformare dati ed eseguire carichi di lavoro. Scopri con quanta rapidità accoglieranno un dashboard che consente loro di risparmiare denaro e distribuire insight nuovi e attendibili.

Riduci il costo del tuo data warehouse fino al 50%[4] grazie all'ottimizzazione dei carichi di lavoro su più motori di query e livelli di storage. Ottimizza i costosi carichi di lavoro del warehouse con motori adatti allo scopo, che aumentano e riducono automaticamente le dimensioni. Riduci i costi eliminando la duplicazione dei dati quando utilizzi storage di oggetti a basso costo; ricava maggior valore da dati in data lake inefficaci.

08

Fasi successive

Sfrutta le conoscenze del team IBM in materia di gestione e ottimizzazione dei dati, affinate da decenni di gestione dei carichi di lavoro più impegnativi al mondo. Scopri con quanta rapidità puoi ottenere valore da watsonx.data.

Entra nella lista d'attesa di watsonx.data →

Contatta il reparto vendite →

1. Why Unstructured Data is the Future of Data Management, Venturebeat, luglio 2021.

2. Worldwide IDC Global DataSphere Forecast, 2022-2026, IDC, maggio 2022.

3. The rise of the data lakehouse: A new era of data value, CIO Magazine, 18 agosto 2022

4. Quando si confrontano i prezzi di listino pubblicati per il 2023, normalizzati per le ore VPC di IBM watsonx.data con alcuni dei principali vendor di data warehouse in cloud. Il risparmio può variare a seconda delle configurazioni, dei carichi di lavoro e dei vendor.

IBM Italia S.p.A.
Circonvallazione Idroscalo
20090 Segrate (Milano)
Italia

Prodotto negli Stati Uniti d'America
Maggio 2023