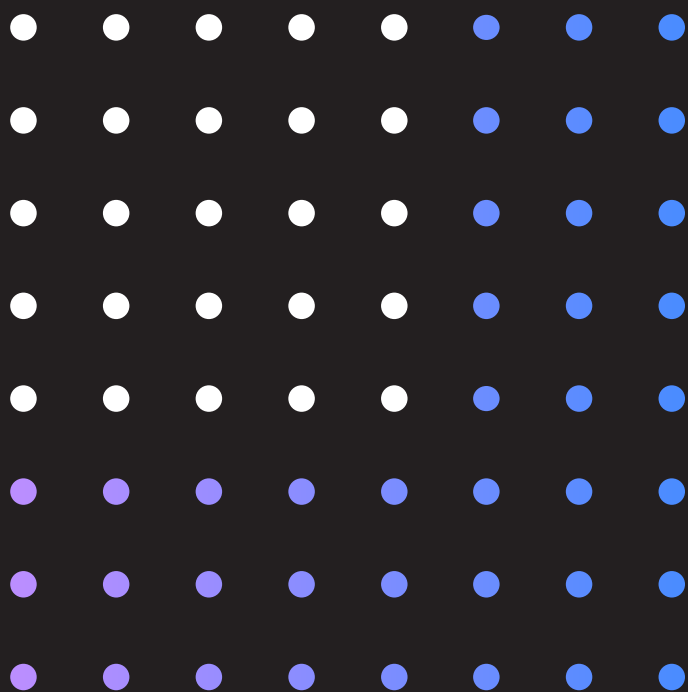


Fornire dati business-ready con catalogazione intelligente dei dati e governance dei data lake

IBM Watson Knowledge Catalog fornisce una piattaforma di governance dei dati basata sull'apprendimento automatico volta ad aiutare ad affrontare le sfide del data lake



Sommario

03

Risolvi le sfide del data lake con un approccio DataOps

03

Sfide per l'utilizzo di data lake aziendali

05

IBM Watson Knowledge Catalog

06

Un'unica fonte di verità e un unico punto di accesso

08

Quattro vantaggi della costruzione di un data lake governato per l'IA

09

Conclusione

Key takeaway

- Poche organizzazioni stanno vedendo il valore che si aspettavano dai data lake che hanno costruito per archiviare e analizzare i loro dati per insight affidabili.
- DataOps risolve le sfide che le organizzazioni affrontano quanto alle inefficienze di accesso, preparazione, integrazione e messa a disposizione dei dati per i consumatori, aderendo alle policy aziendali e normative.
- Le sfide comuni del data lake includono la difficoltà e il costo dell'importazione di nuove source di dati nel data lake, l'incapacità di integrare set di dati interni ed esterni, la mancanza di fiducia nella governance dei dati, l'assenza di accesso agli strumenti di preparazione dei dati self-service e l'impossibilità di trovare e comprendere i dati presenti nel data lake.
- Una piattaforma di governance dei dati aziendali con catalogazione, qualità e rilevamento dei dati può trasformare un deludente progetto di data lake in una vera fonte di valore aziendale.
- [IBM Watson® Knowledge Catalog](#), basato su IBM Cloud Pak™ for Data, fornisce un catalogo di machine learning (ML) per il rilevamento, la catalogazione, la qualità e la governance dei dati. Aiuta gli utenti di dati a scoprire, curare, classificare e condividere rapidamente risorse, set di dati e modelli analitici di dati.
- Quando le organizzazioni mancano di una profonda comprensione dei propri dati, diventa più difficile fidarsi e utilizzare queste informazioni con tutte le forme di intelligenza artificiale (IA), tra cui ML e deep learning.

Risolvi le sfide del data lake con un approccio DataOps

Dieci anni fa, è iniziato il viaggio volto a trovare un approccio flessibile e versatile per costruire un archivio dati centrale in cui far risiedere tutti i dati aziendali. La soluzione era il data lake, un ambiente di archiviazione dei dati per scopi generici che memorizzava praticamente qualsiasi tipo di dati. Consentirebbe inoltre agli analisti aziendali e ai data scientist di applicare i motori e gli strumenti di analisi più appropriati a ciascun set di dati, nella posizione originale.

In genere, questi data lake sono stati creati utilizzando Apache Hadoop e Hadoop Distributed File System (HDFS), combinati con motori come Apache Hive e Apache Spark. Quando questi data lake hanno iniziato a crescere, è emersa una serie di problemi. Mentre la tecnologia era fisicamente in grado di ridimensionare per acquisire, archiviare e analizzare vaste e varie raccolte di dati strutturati e non, è stata prestata troppa poca attenzione alle funzionalità di come incorporare queste capacità nei flussi di lavoro aziendali.

Nel corso del 2022, oltre l'80% dei progetti di data lake non riuscirà a fornire valore poiché la ricerca, l'inventario e la cura dei dati si dimostreranno il principale inibitore del successo di analisi e data science.¹ Di conseguenza, domande come: "Quali dati dovremmo inserire nel data lake?", "Chi li userà?", "Come possiamo facilitarne la ricerca?", "Da dove provengono questi dati?" E "Come possiamo evitare che i dati vengano utilizzati in modo improprio?", spesso è rimasta senza risposta. Questi limiti critici nell'affrontare le persone, i processi e i problemi tecnologici hanno portato effettivamente a far fallire le implementazioni di data lake.

Oggi, molte organizzazioni hanno riconosciuto i loro fallimenti, hanno cambiato i team di leadership per l'implementazione del data lake e stanno lanciando un secondo, terzo o addirittura quarto tentativo volto a implementare correttamente un data lake, questa volta si avvalgono di [DataOps](#) per le operazioni sui dati.

Questo white paper valuta le sfide comuni che devono affrontare i data lake e fornisce nuovi approcci come DataOps che possono aiutare a trasformarli da una palude di dati al centro della pipeline di dati aziendali pronta per l'organizzazione.

DataOps è una pratica di gestione collaborativa dei dati focalizzata su miglioramento della comunicazione, integrazione e automazione dei flussi di dati tra gestori e consumatori di dati all'interno di un'organizzazione.

Presentazione di DataOps

DataOps inserisce le best practice di DevOps, gestione e governance dei dati in un quadro comune, con un modo collaborativo di sviluppare e mantenere flussi di dati tra più parti interessate. DataOps è progettato per risolvere le sfide che le organizzazioni devono affrontare e che sono associate a inefficienze di accesso, preparazione, integrazione e messa a disposizione dei dati per i consumatori, aderendo alle policy

aziendali e normative. Queste efficienze possono essere trovate in una business unit, in un team di analisi o persino in un processo operativo.

Seguire questa metodologia richiede di affrontare le persone, i processi e i problemi tecnologici che fanno la differenza tra implementazioni di data lake efficaci e non. Dal punto di vista tecnologico, DataOps sottolinea l'importanza di utilizzare una piattaforma end-to-end completamente integrata per l'immissione e l'integrazione dei dati, la qualità dei dati, la governance dei dati e il consumo dei dati per creare un data lake governato. Le regole di convalida della qualità dei dati dovrebbero essere eseguite automaticamente come parte del processo di acquisizione per sostenere una pipeline di dati continua in tutta l'azienda. Il processo di ingestione dovrebbe essere completamente integrato con il catalogo di dati, che diventa il fulcro della pipeline. I consumatori di dati dovrebbero essere in grado di accedere ai punteggi di qualità dei dati e ai risultati della profilazione dei dati dal catalogo dei dati e fidarsi dato che l'organizzazione sta lavorando con gli stessi dati nel contesto.

La crescita dei dati sta superando la capacità delle organizzazioni di trarne valore. Quando alle organizzazioni sono state chieste quali sono le maggiori sfide per l'utilizzo dei sistemi di insight hanno risposto: 1) Il 40% sta fondendo i processi aziendali esistenti per reperire dati e analizzarli e 2) Il 39% sta acquistando, acquisendo, gestendo e governando i dati man mano che crescono². Oggi, non si tratta solo di proteggere l'enorme tempo, risorse e investimenti che sono già stati fatti nelle tecnologie del data lake - il fatto è che non esistono alternative. Dall'implementazione dell'IA o persino per eseguire un'analisi completa, è fondamentale avere una visione completa di quanti più dati possibili, il che significa che si ha bisogno di un'architettura in grado di contenere, analizzare e governare tutti quei dati in un unico posto. In molti casi, un data lake governato è l'unica opzione realistica per soddisfare questi requisiti.

Le aziende di oggi possono — e devono — trovare un modo per estrarre valore dal loro data lake assicurando che supporti una pipeline di dati business-ready per DataOps.

Sfide per l'utilizzo di data lake aziendali

Condivisione dei dati

Quando un team all'interno di un'azienda acquisisce o crea un nuovo set di dati, è probabile che abbia un forte senso del valore dei dati e delle sensibilità che li circondano. Se contiene informazioni commercialmente riservate, informazioni di identificazione personale (PII) o dati dei clienti, ad esempio, il team saprà come tali informazioni dovrebbero e non dovrebbero essere utilizzate e adotterà delle precauzioni per assicurarsi che nessuno nel team ne abusi.

Saranno inoltre consapevoli che al di fuori del proprio team, altri potenziali utenti dei dati potrebbero non condividere la stessa comprensione del valore dei dati o dei rischi associati all'uso improprio dei dati. Questi rischi naturalmente renderanno il team estremamente cauto nella condivisione dei dati o nella loro memorizzazione in qualsiasi luogo non sotto il loro controllo.

Questa è una brutta notizia per i data lake. Se l'azienda vede il data lake semplicemente come una discarica incontrollata di dati, sarà molto riluttante ad affidargli i suoi preziosi dati. Di conseguenza, altre parti dell'azienda non saranno in grado di beneficiare di tali dati e l'intero concetto di utilizzare il data lake come repository self-service per la condivisione dei dati aziendali.

Integrazione dei dati

Anche quando un team accetta che i suoi dati vengano integrati nel data lake, questo può essere un processo tortuoso. Il concetto originale del data lake è quello di importare i dati nel loro formato non elaborato, senza richiedere i complessi processi di estrazione, trasformazione e caricamento (ETL) di un tradizionale data warehouse. Tuttavia, la realtà è che quasi tutte le fonti di dati richiedono un certo grado di preelaborazione prima di poter essere utili per qualsiasi tipo di analisi significativa.

Di conseguenza, l'integrazione di una nuova fonte di dati in un data lake può richiedere spesso mesi. Inoltre, poiché gran parte di questi dati è stata precedentemente conservata in piccoli silos operativi anziché in sistemi aziendali, complessivamente, potrebbero esserci dozzine o addirittura centinaia di fonti da integrare.

Ciò significa che in molti casi le informazioni di cui gli analisti aziendali o i data scientist necessitano non sono ancora state aggiunte al data lake e potrebbero non essere aggiunte per mesi o addirittura anni. Ancora una volta, ciò può essere un ostacolo significativo all'adozione.

Archiviazione dei dati

Mentre il costo dello stoccaggio delle materie prime e delle risorse di calcolo è diminuito drasticamente negli ultimi anni, i cluster Hadoop non sono gratuiti. La memorizzazione di enormi quantità di dati in un data lake è molto più conveniente rispetto alla memorizzazione in un'appliance di data warehouse ad alte prestazioni, ma il costo può ancora essere significativo.

Inoltre, a differenza dei dati tradizionalmente archiviati nei data warehouse, il rapporto valore-volume dei big data contenuti in un data lake è relativamente basso.

Se non sai quali set di dati saranno davvero utili e preziosi per i tuoi data scientist, potresti investire ingenti somme nell'integrazione e nell'archiviazione di dati che sono destinati ad affondare nella parte inferiore del tuo data lake e che non verranno mai utilizzati.

Rinvenimento dei dati

Supponendo di aver identificato i set di dati più preziosi da archiviare, di aver convinto gli stakeholder a dividerli

Sfide per l'utilizzo di data lake aziendali



Figura 1. Le imprese che hanno adottato le tecnologie del data lake possono riscontrare una o più di queste problematiche comuni.

e di essere riusciti a integrarli nel data lake, si deve ancora consentire ad altri utenti di trovarli, comprenderli e utilizzarli correttamente. La qualità dei dati nel data lake è un'altra sfida. Non sei sicuro che i dati siano di alta o bassa qualità, ma vengono immessi nel lake.

Sfortunatamente, nella maggior parte dei data lake, ciò non è facile da raggiungere. I dati vengono spesso archiviati senza alcun contesto, rendendo difficile o impossibile per un nuovo utente decodificarli senza consultare il proprietario originale. La terminologia è spesso talmente specifica del dominio che una metrica utilizzata in un'area del business potrebbe essere conosciuta con un nome completamente diverso o definita in modo leggermente diverso da un'altra. Il potenziale di confusione e interpretazione errata può essere così grande che molti set di dati sono effettivamente privi di valore o addirittura pericolosi per un analista che non li conosce già.

Combinazione di dati interni ed esterni

Infine, anche il più grande data lake non dovrebbe tentare di contenere tutti i possibili set di dati che i data scientist della società vorranno utilizzare. Ad esempio, non avrebbe senso importare una replica completa di Google Maps, Weather.com® o Bloomberg nel data lake, solo perché uno dei data scientist desidera eseguire analisi geospaziali o integrare dati meteorologici o prezzi delle azioni in un algoritmo.

Poiché il data lake non deterrà tutti i dati di cui gli analisti aziendali hanno bisogno per l'analisi, dovranno dedicare del tempo a cercarli in più applicazioni. Poiché è probabile

che una percentuale molto elevata di analisi utili implichi la combinazione di set di dati interni ed esterni, ciò aumenta ancora una volta la barriera all'ingresso e dal punto di vista dell'utente, riduce il valore percepito del data lake.

Preparazione dei dati

Esistono molti fattori che rendono difficile [la preparazione dei dati](#), dalla comprensione della posizione in cui trovare i dati alla formattazione. La preparazione dei dati ai fini del loro uso nell'analisi è l'attività più inefficiente e che richiede tempo per gli utenti di dati. Gli utenti di dati dedicano la maggior parte del loro tempo a trovare, pulire e formattare le informazioni, invece di concentrarsi sull'analisi dei dati, sulla modellazione e sulla comprensione delle informazioni per l'impatto aziendale.

L'accessibilità limitata ai set di dati governati ha inoltre causato una dipendenza eccessiva dall'IT durante la fase di preparazione. Questo accesso limitato segnala la necessità di migliorare le capacità self-service e le competenze di alfabetizzazione dei dati in tutta l'azienda per alleviare questo blocco.

Qualità dei dati

Il dumping dei dati in un data lake può renderli inutilizzabili. Poiché non esistono regole sulla qualità dei dati o sulla convalida da applicare ai dati prima che vengano inseriti in un data lake, non vengono forniti dati che possono essere considerati affidabili e utilizzati. Dati di alta qualità sono una caratteristica essenziale che determina l'affidabilità dei dati per prendere decisioni. I dati sono un bene prezioso che deve essere gestito mentre si si muove all'interno di un'organizzazione. Poiché le fonti di informazione stanno diventando sempre più numerose e diversificate e le iniziative di conformità normativa più focalizzate, la necessità di integrare e accedere alle informazioni da queste fonti disperate in modi coerenti, affidabili e riutilizzabili è fondamentale.

Un approccio olistico alla costruzione di data lake governati

La maggior parte dei data lake sfrutta Apache Hadoop e il suo ampio ecosistema di progetti open source per i loro livelli di archiviazione dei dati e motori di analisi. Non sorprende che la comunità open source che gira intorno a Hadoop abbia riconosciuto i problemi che devono affrontare le attuali implementazioni del data lake e recentemente sono sorti molti progetti, che mirano a risolvere i vari problemi individualmente. Allo stesso modo, ci sono una serie di strumenti proprietari sul mercato che pretendono di risolvere gli stessi problemi.

Potrebbe essere allettante, quindi, rimediare ai problemi dei dati in modo frammentario, non appena si presentano. Quando il numero di set di dati aumenta troppo per essere gestibile, si deve aggiungere uno strumento di catalogazione. Quando gli utenti si lamentano di non poter trovare i dati di cui hanno bisogno, restano imbullonati a un front-end con una funzione di ricerca. E quando gli amministratori di dati non possono più tenere traccia della provenienza dei dati o di chi li utilizza, si deve implementare strumenti di derivazione dei dati e un framework di governance dei dati.

Sembra semplice, ma, in pratica, questo approccio frammentario tende a comportare un costo per aumentare enormemente la complessità e ridurre la manutenibilità, soprattutto con l'aumentare della scala e della portata del

data lake. Allo stesso modo in cui l'aggiunta di nuove origini dati a un data lake aumenta la complessità dei requisiti ETL, l'aggiunta di nuovi strumenti tende ad aumentare la complessità dei requisiti non funzionali del data lake.

Invece di avere una piattaforma end-to-end integrata in grado di integrare i dati, si deve eseguire operazioni di qualità sui dati e catalogare i dati per un uso efficace da parte degli analisti aziendali per, in genere, scoprire che ogni strumento ha i suoi modi di gestire i guasti e il proprio approccio alla registrazione. Di conseguenza, la risoluzione dei problemi può richiedere molto tempo.

Un altro, più importante difetto dell'approccio frammentario diventa evidente quando si assume una visione meno tecnica e più concettuale dei problemi che comunemente incontrano i data lake. L'intuizione chiave è che scalabilità, reperibilità, integrazione, qualità dei dati e governance non sono problemi separati: sono indissolubilmente correlati. Risolverli richiederà un approccio molto più olistico.

Scalabilità, reperibilità, integrazione, qualità dei dati e governance non sono problemi separati: sono indissolubilmente correlati. Risolverli richiederà un approccio olistico alla gestione delle informazioni.

IBM Watson Knowledge Catalog Rilevazione dei dati, catalogazione dei dati e qualità dei dati

L'[IBM Watson Knowledge Catalog](#) basato su IBM Cloud Pak for Data aiuta gli utenti di dati a scoprire, curare, classificare e condividere rapidamente risorse di dati, set di dati, modelli analitici e le loro relazioni con altri membri dell'organizzazione. Aiuta i team di governance dei dati a definire il glossario, le politiche e le regole aziendali e fornisce flussi di lavoro avanzati per la governance. Il catalogo funge da unica fonte di verità per i data engineer, amministratori dei dati, datascientist e analisti aziendali per ottenere l'accesso self-service ai dati su cui confidare e da utilizzare con fiducia.

Soluzioni come IBM Watson Knowledge Catalog fornito da IBM Cloud Pak for Data possono fornire tutte le funzionalità necessarie per risolvere i principali problemi dei data lake di oggi in un'unica piattaforma completa. Il catalogo aiuta a risolvere la causa principale di questi problemi correlati: il diffuso insuccesso dei data lake nel fornire strumenti efficaci per acquisire, archiviare e gestire i metadati e tenere traccia dell'origine dei dati.

In molti modi, il valore di un data lake dipende dai metadati che contiene, tanto quanto dipende dai dati stessi. Senza metadati che spiegano da dove proviene un set di dati, chi l'ha creato, cosa contiene, chi è autorizzato a usarlo e come viene utilizzato, i dati stessi sono praticamente privi di valore. Gli utenti non saranno in grado di trovarli e, anche se lo fanno, non capiranno cosa significa, non si fideranno né sapranno come utilizzarli.

Watson Knowledge Catalog

Fornisci dati attendibili e significativi

Organizza i tuoi dati



Conosci

I dati devono essere completi, applicabili e accessibili ovunque. Scopri, classifica e capisci tutti i tipi di dati.

Governa i tuoi dati



Abbi fiducia

I dati devono essere sicuri, puliti e facili da trovare per incoraggiare l'accesso self-service di fiducia. Comprendi da dove provengono i dati e la loro qualità.

Democratizza i tuoi dati



Consuma

Capacità di guidare la scoperta self-service e automatizzare il processo decisionale per far evolvere il business. Fornisci una visione di tutte le informazioni a coloro che ne hanno bisogno e consenti loro di accedervi.

Figura 2. IBM Watson Knowledge Catalog offre una vasta gamma di funzionalità per il rilevamento, la catalogazione e la governance dei dati.

Un'unica fonte di verità e un unico punto di accesso

IBM Watson Knowledge Catalog basato su IBM Cloud Pak for Data risolve questi problemi rendendo i metadati una priorità fondamentale. Al centro c'è un potente motore di catalogazione che indicizza tutti i set di dati e le risorse analitiche a cui la tua azienda ha accesso, indipendentemente da dove risiedono i tuoi dati, nel tuo data lake, data warehouse o sistema transazionale, o anche in una serie di fogli di calcolo. Indipendentemente dal fatto che siano strutturati o meno o archiviati on premise o ospitati nel cloud. Inoltre, il catalogo può includere anche set di dati esterni e fonti, come servizi dati proprietari a cui la tua azienda si abbona o API open data.

Oltre a fornire un'unica fonte di verità su tutti i tuoi set di dati, il catalogo dati fornisce anche un unico punto di accesso. La ricerca basata sull'intelligenza artificiale e le funzionalità di suggerimento aiutano gli analisti aziendali, i data scientist, i tecnici della qualità dei dati e i team di governance dei dati a trovare le risorse più facilmente e presentano i metadati disponibili per aiutare gli utenti a capire cosa hanno trovato e valutare se è utile.

Le funzionalità integrate di preparazione dei dati self-service accelerano il tempo necessario per trasformare i dati per un uso produttivo in applicazioni di analisi e IA. Gli analisti aziendali e i data scientist non devono perdere tempo a preparare e analizzare i dati. L'integrazione con una soluzione di preparazione dei dati a livello aziendale, come [IBM® InfoSphere® Advanced Data Preparation](#), contribuisce a garantire che i set di dati governati creati attraverso la superficie del catalogo a quelli con il maggior contesto per favorire insight e azioni aziendali per gli utenti aziendali. Questa integrazione favorisce la collaborazione attraverso la pipeline di dati.

Scalabilità, reperibilità, integrazione, qualità dei dati e governance non sono problemi separati: sono indissolubilmente correlati. Risolverli richiederà un approccio olistico alla gestione delle informazioni.

Il catalogo aiuta anche gli amministratori dei dati che operano presso l'ufficio del Chief Data Officer (CDO) codificando e classificando i set di dati e tracciando automaticamente la loro discendenza e il loro utilizzo e sfruttando il glossario aziendale incorporato per standardizzare la terminologia aziendale attraverso i dati. Di conseguenza, è più facile per gli amministratori capire cosa contiene ogni set di dati, dove si trovano le informazioni sensibili o PII e chi dovrebbe essere autorizzato ad accedervi.

Un unico catalogo per più origini dati all'interno e all'esterno dell'organizzazione

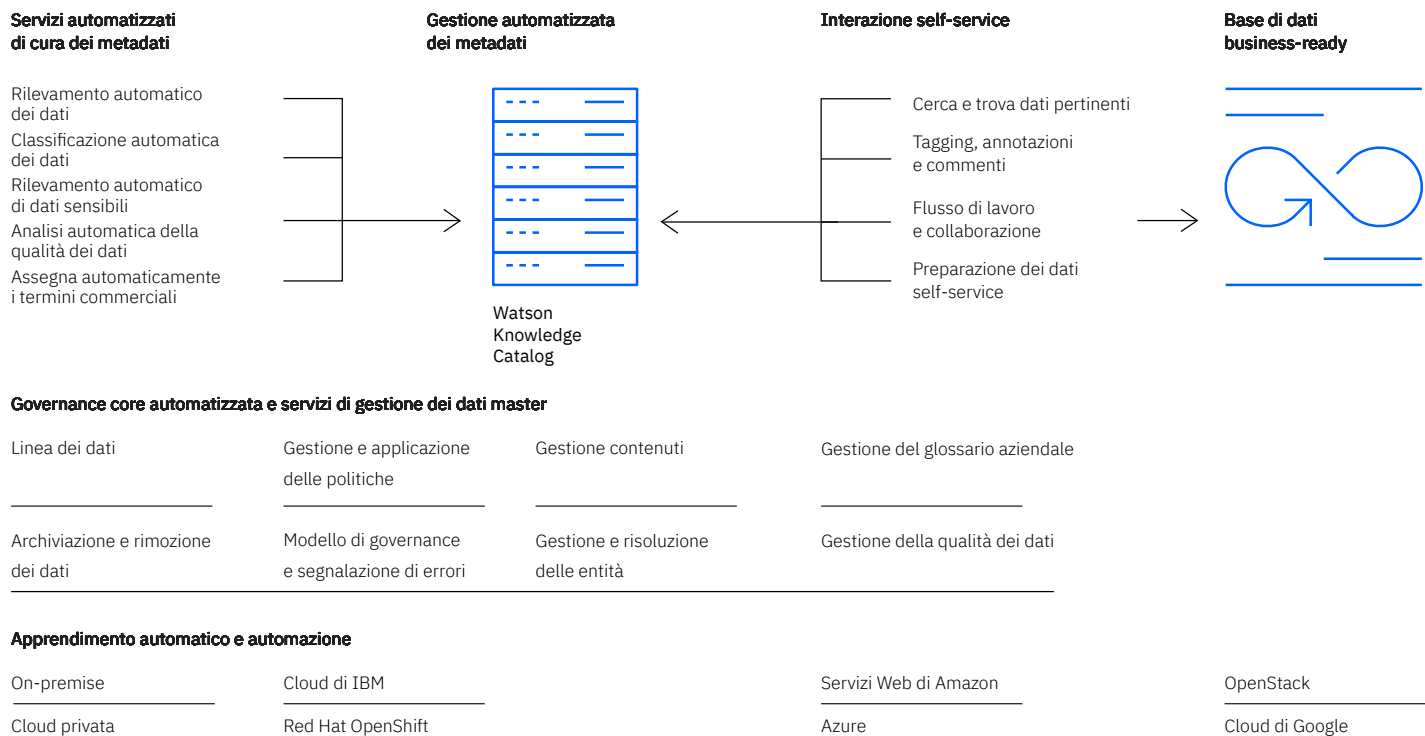


Figura 3. Con l'indice di metadati intelligente di IBM Watson Knowledge Catalog, i dati, sia strutturati che non, possono risiedere nei sistemi originali, ma gli utenti possono scoprirli rapidamente per un'analisi più intelligente.

IBM Watson Knowledge Catalog rende i metadati una priorità fondamentale, fornendo un'unica fonte di verità e un unico punto di accesso a tutti i set di dati a cui la tua azienda ha accesso.

Rilevamento intelligente integrato dei dati

Per migliorare ulteriormente la reperibilità, il catalogo consente agli utenti di taggare e commentare set di dati e risorse analitiche, arricchendo i metadati e aggiungendo ulteriore contesto per aiutare i colleghi a trovare ciò di cui hanno bisogno. La soluzione include anche algoritmi di rilevazione dei dati integrati che utilizzano ML per classificare automaticamente i contenuti di ciascun set di dati. Identificando i tipi di campi comuni come nomi, indirizzi, codici postali e numeri di previdenza sociale, la soluzione riduce la necessità per gli autori di annotare manualmente i dati. Integra automazione e ML per automatizzare la gestione dei dati e la gestione dei metadati. Con funzioni integrate di qualità dei dati, la soluzione consente di profilare in profondità i dati, la qualità dei dati e le regole di convalida.

Le operazioni automatizzate sui dati forniscono una pipeline di dati curata, con qualità e governance dei dati e contribuiscono a garantire un flusso continuo di dati governati di alta qualità nel data lake.

Allo stesso modo, l'aggiunta di un modello di metadati intelligente delle risorse fornisce un modo unico per applicare automaticamente come il Regolamento generale sulla protezione dei dati (GDPR) e il California Consumer Privacy Act (CCPA).

IBM Watson Knowledge Catalog basato su IBM Cloud Pak for Data contribuisce a fornire dati affidabili, di alta qualità e pronti per il business a praticamente tutti gli utenti di dati.

Tutti i componenti della soluzione sono stati progettati come microservizi, con un unico insieme di principi di progettazione e un approccio comune ai requisiti non funzionali, come scalabilità, gestione degli errori, sicurezza e registrazione.

IBM Watson Knowledge Catalog fornisce una piattaforma di governance aziendale ML, quindi è pronta per l'IA su vasta scala.

Invece degli errori confusi e dei colli di bottiglia delle prestazioni che potrebbero derivare da un approccio frammentario e fai-da-te, IBM Watson Knowledge Catalog fornisce una piattaforma di governance aziendale ML, quindi pronta per l'IA su vasta scala.

IBM Watson Knowledge Catalog è disponibile in tre varianti:

- Come soluzione SaaS (Software as a service) in IBM Cloud™
- In [IBM Cloud Pak for Data](#)
- Integrato con [IBM Watson Studio](#)

Soluzioni come IBM Watson Knowledge Catalog possono sbloccare il valore che le iniziative del data lake promettevano originariamente. Watson Knowledge Catalog con capacità di catalogazione e governance intelligenti aiuta a costruire un data lake affidabile e governato per l'IA.

Quattro vantaggi della costruzione di un data lake governato per l'IA

1. Crea la fiducia nei dati attraverso la qualità e la governance

- Le funzionalità di qualità dei dati aiutano a migliorare la qualità dei dati e a renderli disponibili nel data lake.
- I criteri di governance vengono impostati e applicati automaticamente, quindi, quando si trova un set di dati, si sa se e come è permesso utilizzarlo.
- Puoi curare i tuoi dati man mano che gli utenti aggiungono valutazioni, commenti e altre informazioni che aiuteranno gli altri a determinare se un set di dati sarà utile.

2. Autorizza i tuoi utenti di dati

- I tuoi team line-of-business (LOB) condividono i loro dati volentieri perché sono fiduciosi che saranno adeguatamente governati e protetti da abusi.
- Puoi favorire la collaborazione e trasformare i dati in risorse aziendali affidabili attraverso policy e applicazione dinamiche dei dati.
- I tuoi dati diventano più reperibili e riutilizzabili nel tempo, poiché gli utenti aggiungono tag e metadati pertinenti per aiutare gli altri a estrarne valore.
- Una singola interfaccia ti dà accesso a tutti i set di dati che la tua organizzazione possiede, indipendentemente da dove siano archiviati.

3. Torna indietro

- Il rilevamento automatico dei dati riduce il tempo e lo sforzo necessari per aggiungere metadati per nuovi set di dati.
- La gestione automatica dei dati e la gestione dei metadati riducono il tempo necessario per rilevare i metadati e assegnare i termini, nonché i tempi di creazione del glossario aziendale.

- Con strumenti di preparazione dei dati self-service semplici e intuitivi, i tuoi utenti dedicano meno tempo a preparare i dati e più tempo a scoprire insight.
- Scateni i tuoi data scientist e i tuoi analisti aziendali per fornire analisi migliori in un breve lasso temporale.
- La ricerca intelligente e basata sull'intelligenza artificiale ti aiuta a trovare i dati di cui hai bisogno in pochi secondi, invece di aspettare settimane affinché un altro team li fornisca.

4. Gestisci dati e costi in crescita

- È possibile ottimizzare i costi di archiviazione evitando le spese di inserimento di set di dati di basso valore nel data lake.
- È possibile anche vedere tutti i set di dati esterni a cui la propria organizzazione si abbona, riducendo il rischio di pagare più abbonamenti di quelli di cui si ha bisogno.
- È possibile dare la priorità all'ingestione di nuove fonti di dati nel data lake in base alla richiesta dei dati da parte degli utenti, contribuendo a integrare prima le fonti più preziose.

Sblocca il valore dei tuoi dati

Sia che lavori nell'ufficio del CDO, nel dipartimento IT o come scienziato o analista di dati LOB, tu e i tuoi colleghi condividete un obiettivo comune. Se riesci a costruire un data lake che mantenga davvero le promesse, non solo potresti rendere i tuoi task molto più facili e produttivi. Inoltre, potresti svolgere un ruolo chiave nel fornire alla tua azienda un vantaggio competitivo con il quale poche organizzazioni attualmente possono competere.

Se riesci a ripulire il tuo data lake mentre i tuoi concorrenti continuano a vacillare nel caos, scoprirai delle possibilità che solo pochi possono aspettarsi. L'autentico vantaggio della prima mossa attende coloro che sono i primi a sbloccare il valore dei dati precedentemente non sfruttati.

Conclusione

Scopri dove risiedono tutti i tuoi dati, chi li sta usando e il suo valore per la tua attività di analisi.

Le iniziative critiche per DataOps sono i cataloghi di dati perché possono contribuire a fornire una gestione automatizzata dei metadati aperti integrando governance dei dati, qualità e gestione attiva delle policy.

IBM Watson Knowledge Catalog con capacità di catalogazione e governance intelligenti aiuta a costruire un data lake affidabile e governato per l'IA. Il catalogo integra l'integrazione dei dati, la qualità e la governance dei dati nell'ambiente del data lake per aiutare a fornire dati business-ready per DataOps e un'unica fonte di verità.

Per ulteriori informazioni

Per saperne di più, visita:

ibm.com/cloud/watson-knowledge-catalog

© Copyright IBM Corporation 2019

IBM Italia S.p.A.

Circonvallazione Idroscalo
20090 Segrate (MI)
Italia

Prodotti negli Stati Uniti d'America ad ottobre 2019 IBM, il logo IBM, ibm.com, IBM Cloud, IBM Cloud Pak, IBM Watson e InfoSphere sono marchi di International Business Machines Corp., registrati in molte giurisdizioni in tutto il mondo.

Red Hat e OpenShift sono marchi commerciali o marchi registrati di Red Hat, Inc. o delle sue consociate negli Stati Uniti e in altri Paesi. I nomi di altri prodotti e servizi possono essere marchi di IBM o di altre aziende. Un elenco attuale dei marchi IBM è disponibile sul Web in "Informazioni sul copyright e sui marchi" all'indirizzo www.ibm.com/legal/copytrade.shtml

Il presente documento è aggiornato alla data di pubblicazione iniziale e può essere modificato da IBM in qualsiasi momento. Non tutti i prodotti e servizi sono disponibili in tutti i paesi in cui IBM opera. L'informazione è che questo documento è fornito "così com'è" senza alcuna garanzia, espressa o implicita, incluso senza alcuna garanzia di commerciabilità, idoneità per uno scopo particolare e alcuna garanzia o condizione di non violazione. I prodotti IBM sono garantiti in base ai termini e alle condizioni degli accordi in base ai quali vengono forniti. Il cliente è responsabile di garantire la conformità alle leggi e ai regolamenti applicabili. IBM non fornisce consulenza in materia legale, né dichiara o garantisce che i propri prodotti e servizi assicurino che il cliente sia in conformità a qualunque disposizione di legge.

1. Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders—Gartner, settembre 2019
2. The Forrester Wave: Machine Learning Data Catalogs, Q2 2018

ASW12449-ITIT-03

