Qlik

WHITE PAPER

QUALITÀ E GOVERNANCE DEI DATI

I sei criteri dei dati pronti per l'IA

Stabilire una base di dati solida per l'IA

Indice

Sintesi

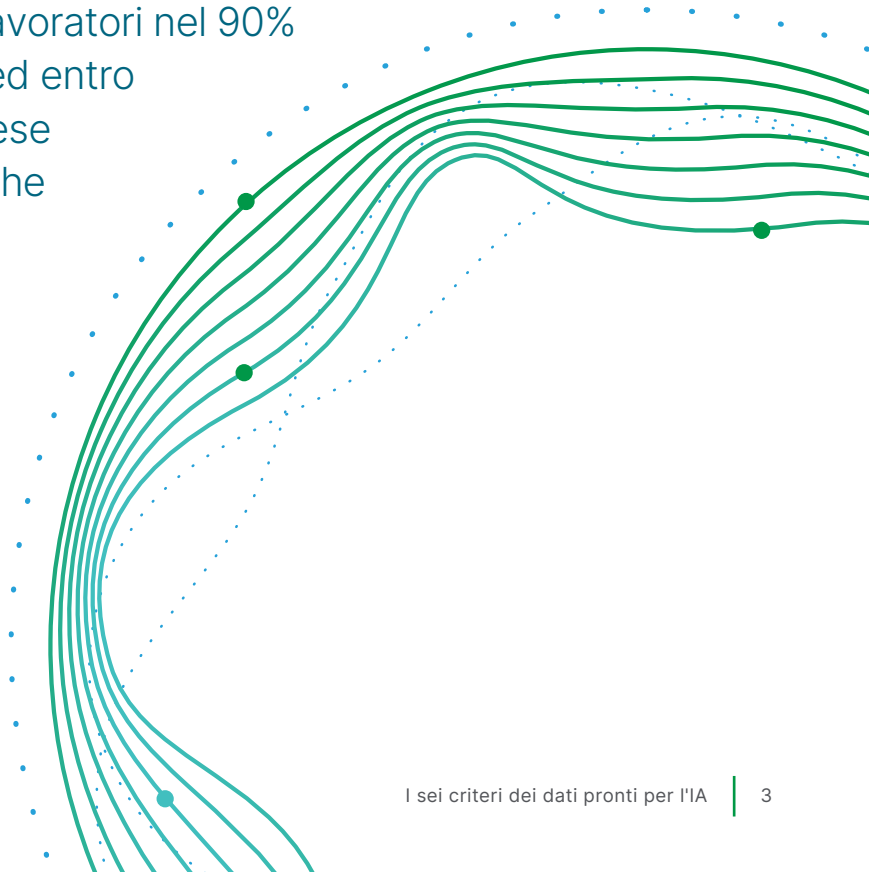
- Questo documento individua sei criteri per verificare che i dati siano pronti per essere utilizzati dall'intelligenza artificiale (IA).
- I sei criteri prevedono che i dati siano diversificati, aggiornati, accurati, sicuri, rilevabili e facilmente fruibili dai sistemi.
- Inoltre, il presente documento descrive AI Trust Score, uno strumento che aiuta a determinare in che misura i dati dell'azienda rispettano tali criteri.

Introduzione

Si ritiene che l'intelligenza artificiale (IA) porterà grandi innovazioni in settori come la sanità, il manifatturiero e il customer service, favorendo esperienze di alta qualità per clienti e dipendenti. Le tecnologie basate sull'IA come il machine learning (ML) hanno già aiutato i professionisti dei dati a produrre previsioni matematiche, generare insight e prendere decisioni migliori. Inoltre, le nuove tecnologie come l'IA generativa (GenAI) sono in grado di creare contenuti altamente realistici che possono potenzialmente aumentare la produttività in qualsiasi area dell'azienda.

Secondo Gartner, entro il 2025 l'IA generativa affiancherà i lavoratori nel 90% delle aziende multinazionali ed entro il 2026 oltre l'80% delle imprese avrà impiegato applicazioni che sfruttano l'IA generativa nella produzione[1].

Tuttavia, l'IA funziona bene solo con dati di qualità e questo white paper descrive i sei criteri per verificare che i dati siano pronti per l'IA.



[1] "Analysts to Discuss Generative AI Trends and Technologies," Gartner, ottobre 2023.

I sei criteri dei dati pronti per l'IA

Sarebbe sciocco pensare di poter semplicemente inserire i dati nelle varie iniziative di IA e aspettarsi che succeda qualcosa, eppure è proprio quello che fanno molti professionisti. Anche se questo approccio sembra funzionare nei primi progetti, via via che questi si espandono i data scientist impiegano molto più tempo a correggere e preparare i dati.

Inoltre, i dati utilizzati per l'IA devono essere di alta qualità e preparati appositamente per tali applicazioni. Questo comporta diverse ore di lavoro manuale per ripulire e migliorare i dati, in modo da assicurare precisione e completezza e un'organizzazione che sia facilmente comprensibile dalle macchine. Inoltre, questi dati richiedono spesso informazioni aggiuntive, come definizioni ed etichette, per arricchire il significato semantico necessario all'apprendimento automatico e per consentire all'IA di operare in modo più efficiente.

Dunque, prima i dati vengono preparati per i processi di IA a valle, maggiore è il beneficio. Usare dati preparati e pronti per l'IA è come dare a uno chef verdure già lavate e tagliate invece che consegnargli direttamente la spesa: riduce tempi e fatica e consente di preparare il piatto con celerità. Il diagramma qui sotto mostra i sei criteri fondamentali per assicurare la "prontezza" dei dati e la loro sostenibilità per l'utilizzo da parte dell'IA.

Le sezioni seguenti descrivono in dettaglio questi criteri.

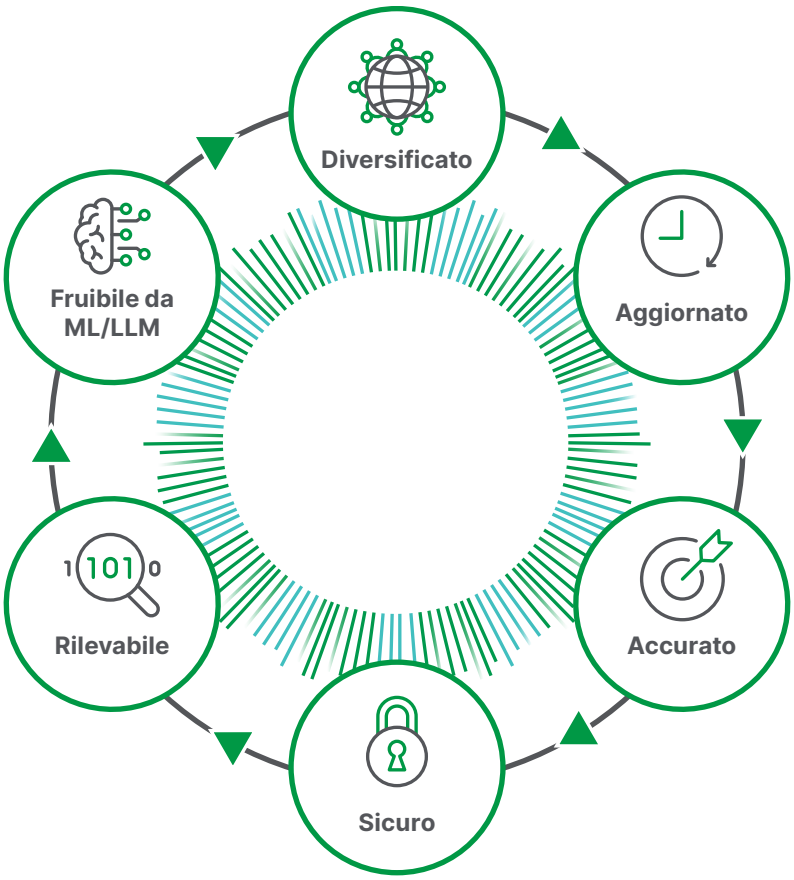


Figura 1. I sei criteri dei dati pronti per l'IA

1 I dati devono essere diversificati.

I pregiudizi nei sistemi di IA, detti anche machine learning bias o bias degli algoritmi, si incontrano quando le applicazioni di IA producono risultati che riflettono i pregiudizi umani, come le ineguaglianze sociali. Ciò accade quando il processo di sviluppo di un algoritmo include supposizioni pregiudiziali o, più frequentemente, quando i dati di addestramento contengono bias. Per esempio, un algoritmo di affidabilità creditiza potrebbe negare un prestito se si basa sempre e solo su un ventaglio ristretto di attributi economici.

Di conseguenza, il nostro primo criterio suggerisce di fornire un'ampia varietà di dati ai modelli di IA, per aumentarne la diversità, ridurre i bias e assicurarsi che le applicazioni di IA siano meno propense a prendere decisioni ingiuste.

Avere dati diversificati significa costruire i modelli di IA senza basarsi su set di dati ristretti e in silos. Al contrario, i dati vanno estratti da una molteplicità di sorgenti con diversi pattern, prospettive, variazioni e scenari rilevanti per il campo di applicazione. Questi dati possono essere ben strutturati e trovarsi nel cloud o on-premise. Possono anche trovarsi in mainframe, database, sistemi SA o applicazioni SaaS (software as a service). Al contrario, i dati sorgente possono essere non strutturati e trovarsi in file o documenti di un drive aziendale.

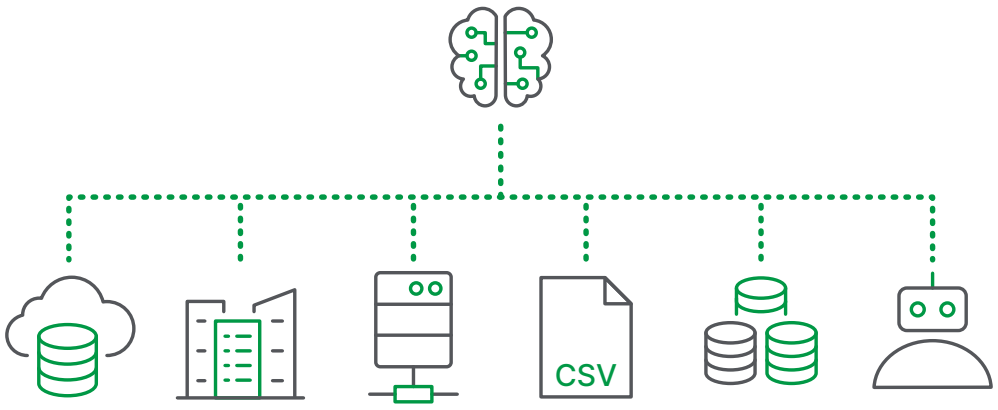


Figura 2. Diversificazione dei dati

È fondamentale acquisire dati diversificati in varie forme, da integrare nelle applicazioni di ML e GenAI.

2 I dati devono essere aggiornati.

È vero che le applicazioni di ML e GenAI hanno bisogno di dati diversificati, ma è importante anche che siano recenti. Come una previsione meteo basata sulle condizioni meteorologiche del giorno precedente non è utile per organizzare un'escursione, i modelli addestrati dall'IA con informazioni non aggiornate possono produrre risultati inaccurati o irrilevanti. Inoltre, i dati più recenti consentono ai modelli di IA di essere sempre aggiornati sui trend, adattarsi a circostanze mutevoli e fornire migliori risultati. Dunque, il secondo criterio dei dati pronti per l'IA è la tempestività.

Per i progetti di IA, è fondamentale sviluppare e utilizzare pipeline a bassa latenza e in tempo reale che garantiscano l'uso di dati aggiornati. **La tecnologia change data capture (CDC)** è spesso utilizzata per fornire dati aggiornati da sistemi di database relazionali e la **stream capture** è utilizzata per i dati provenienti da dispositivi IoT che necessitano di un'elaborazione a bassa latenza. Una volta acquisiti i dati, i repository target vengono aggiornati e **modificati in modo costante**, quasi in tempo reale, per avere dati il più aggiornati possibile.

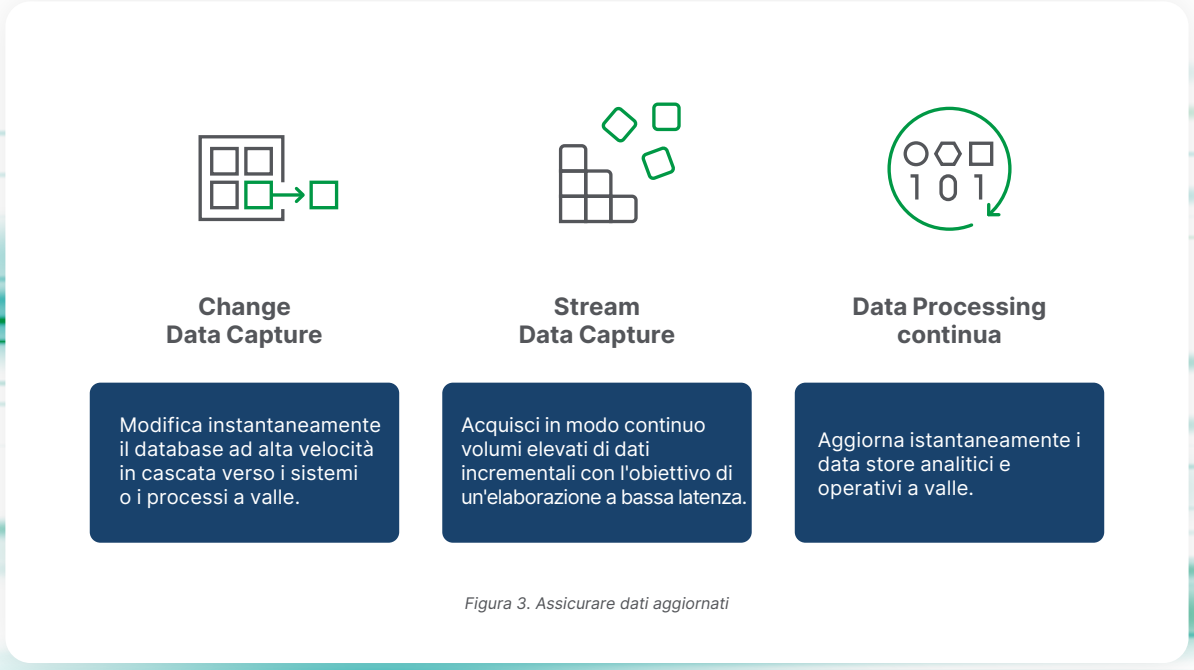


Figura 3. Assicurare dati aggiornati

Come abbiamo detto, dati aggiornati consentono di produrre previsioni più accurate e consapevoli.

3 I dati devono essere accurati.

Il successo di qualsiasi iniziativa di ML o GenAI si basa su un elemento fondamentale: la correttezza dei dati. Questo perché i modelli di AI sono come sofisticate spugne che assorbono le informazioni per apprendere ed eseguire compiti. Se un'informazione non è accurata, è come se la spugna assorbisse acqua sporca, restituendo risultati basati su bias, privi di senso e, in ultima analisi, compromettendo il funzionamento del sistema di IA. Dunque l'accuratezza dei dati è il terzo criterio ed è un elemento fondamentale per creare applicazioni di IA affidabili.

L'accuratezza dei dati ha tre aspetti. Il primo è la **profilazione dei dati sorgente** per comprenderne le caratteristiche, la completezza, la distribuzione, la ridondanza e la forma. La profilazione è anche comunemente nota come analisi esplorativa dei dati o EDA (exploratory data analysis).

Il secondo aspetto include le **strategie di riduzione dell'operazionalizzazione** attuate costruendo, utilizzando e monitorando costantemente l'efficacia delle regole di qualità dei dati. In questa fase si devono coinvolgere i data steward per contribuire alla deduplicazione e all'unificazione dei dati. In alternativa, l'IA può aiutare ad automatizzare e accelerare il processo fornendo suggerimenti di data quality.

L'ultimo aspetto riguarda la provenienza dei dati e le analisi di impatto, da svolgersi con strumenti per scienziati e ingegneri di dati che rilevino l'impatto di potenziali variazioni e traccino l'origine dei dati per prevenire modifiche accidentali dei dati usati dai modelli di IA.

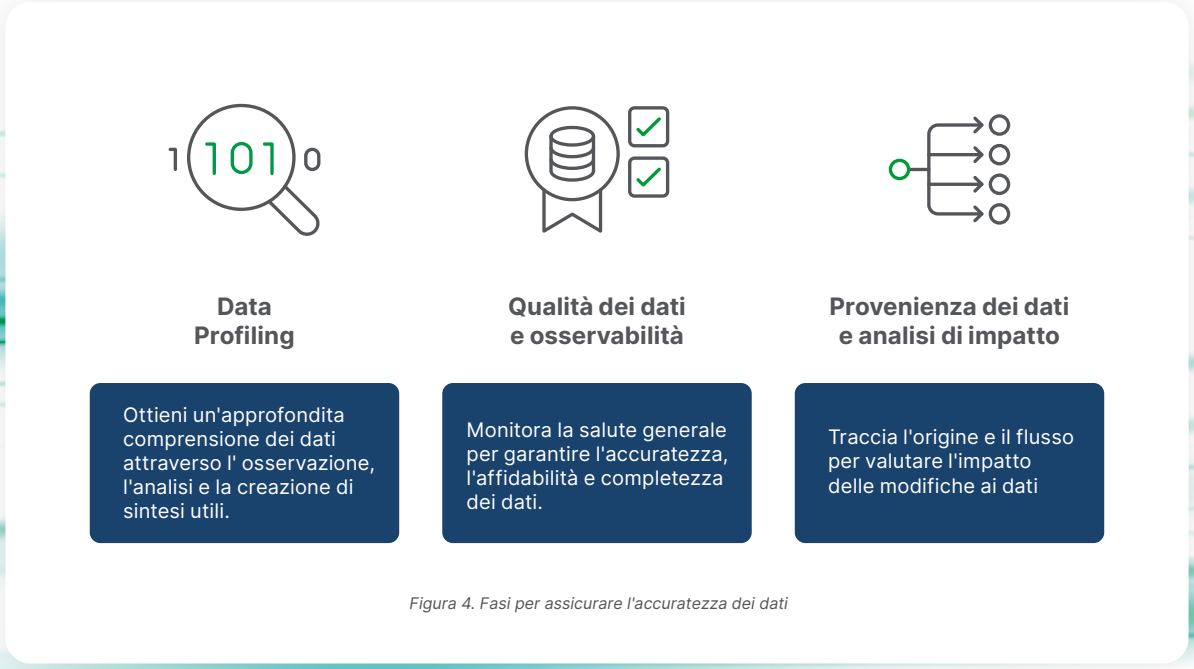


Figura 4. Fasi per assicurare l'accuratezza dei dati

Dati accurati e di qualità consentono ai modelli di identificare pattern e relazioni importanti, portando a decisioni, generazioni e previsioni più precise.

4 I dati devono essere sicuri

Spesso i sistemi di IA usano dati sensibili, come informazioni sull'identità (PII), dati finanziari o informazioni protette da copyright, che richiedono responsabilità. Lasciare i dati non protetti in applicazioni di IA è come lasciare aperta la porta di una cassaforte. Qualche malintenzionato potrebbe rubare informazioni sensibili, manipolare i dati di addestramento portando a risultati basati su bias o addirittura distruggere l'intero sistema di GenAI. Mettere al sicuro i dati è fondamentale per proteggere la privacy, mantenere integro il modello e assicurare uno sviluppo responsabile di potenti applicazioni di IA. Per queste ragioni, la sicurezza è il quarto criterio dei dati pronti per l'IA.

Anche in questo caso esistono tre strategie che aiutano ad automatizzare la sicurezza su larga scala, dato che è quasi impossibile farlo manualmente. **La classificazione** rileva, classifica ed etichetta i dati che passano al livello successivo. **La protezione** definisce le policy come schermatura, tokenizzazione e crittografia per offuscare i dati. Infine, la **sicurezza** definisce le linee guida per il controllo degli accessi ai dati. Questi tre concetti si integrano come segue: in primo luogo si definiscono i livelli di privacy e i dati vengono etichettati con come *sensibili*, *confidenziali* o *riservati*. Poi viene applicata una policy di protezione per mascherare i dati riservati. Infine, viene utilizzata una policy per gestire il diritto di accesso ai dati.

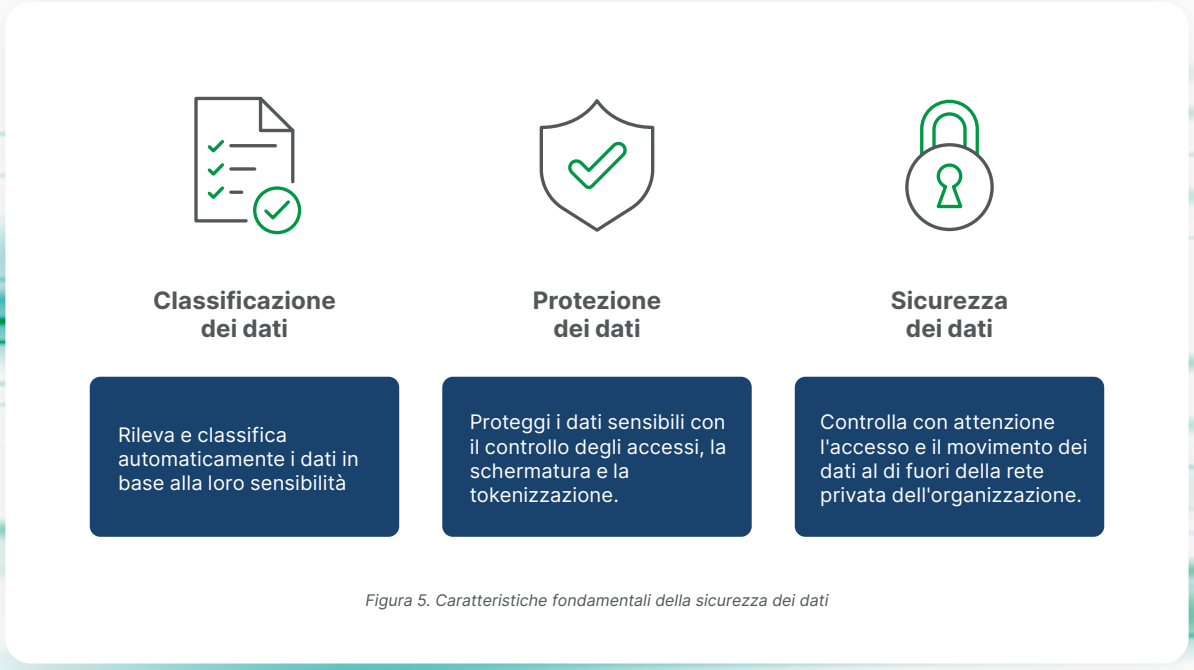


Figura 5. Caratteristiche fondamentali della sicurezza dei dati

Queste tre strategie proteggono i dati e sono fondamentali per aumentare la fiducia complessiva nei sistemi di IA e salvaguardarne la reputazione.

5 I dati devono essere rilevabili.

I criteri che abbiamo visto finora si concentrano soprattutto sulla fornitura tempestiva di dati corretti, nel giusto formato, a determinati sistemi, persone o applicazioni di IA. Ma accumulare dati non è sufficiente. I dati pronti per l'IA devono essere individuabili e facilmente rilevabili all'interno del sistema. Altrimenti sarebbe come avere una libreria con tutti i testi sottochiave: le informazioni sono lì, ma non sono utilizzabili. I dati che possono essere rilevati sbloccano il vero potenziale di ML e GenAI, consentendo ai carichi di lavoro di trovare le informazioni necessarie per apprendere, adattarsi e produrre risultati d'impatto. Per questo l'accessibilità è il quinto criterio dei dati pronti per l'IA.

Non sorprende dunque che le buone pratiche relative ai metadata si basino sull'accessibilità. Oltre ai metadati tecnici associati ai set di dati di IA, vanno definiti anche i metadati commerciali e la tipizzazione semantica. La **tipizzazione semantica** fornisce un significato ulteriore ai sistemi automatizzati, mentre i termini commerciali addizionali forniscono maggiore contesto per facilitare la comprensione umana. È bene creare un **glossario aziendale** che associ i termini commerciali alle entità tecniche nei set di dati, per garantire una comprensione comune dei concetti. Si può anche utilizzare l'incremento basato sull'IA per generare automaticamente documenti e aggiungere semantiche commerciali dal glossario. Così tutti i metadati vengono indicizzati e resi consultabili tramite un **catalogo di metadati.**

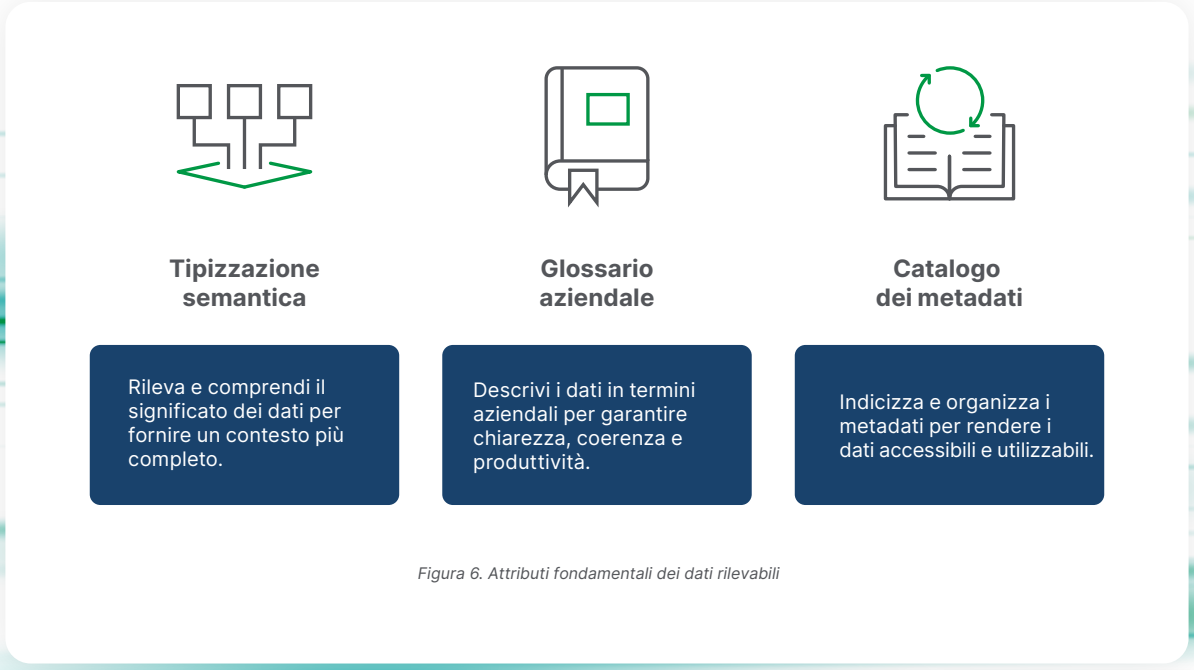


Figura 6. Attributi fondamentali dei dati rilevabili

Questo approccio garantisce che I dati siano direttamente e facilmente rilevabili consultabili, applicabili, utilizzabili e validi per l'IA.

6 I dati devono essere facilmente fruibili per ML o LLM.

Abbiamo già detto che le applicazioni di ML e GenAI sono strumenti potenti, ma il loro potenziale dipende dalla possibilità di accedere rapidamente ai dati. A differenza delle persone, che possono decifrare appunti scritti a mano o muoversi tra tabelle disordinate, queste tecnologie hanno bisogno di informazioni rappresentate in formati specifici. È come servire un pasto a una persona schizzinosa: se non mangia quel che le viene servito, resterà con la fame. Allo stesso modo, i progetti di IA non hanno successo se i dati non sono nel formato giusto per esperimenti di ML o applicazioni di LLM. Rendere i dati facilmente fruibili aiuta a sfruttare il potenziale di questi sistemi di IA, permettendo una facile acquisizione delle informazioni, che vengono tradotte in azioni intelligenti per fornire risultati creativi. Di conseguenza, assicurare una rapida fruizione dei dati è il sesto e ultimo criterio dei dati pronti per l'IA.

Rendere i dati fruibili per il machine learning.

La trasformazione dei dati è il segreto per avere dati fruibili per il ML. Gli algoritmi come le regressioni lineari rubano la scena, ma la qualità e la forma dei dati su cui sono addestrati sono altrettanto importanti. Inoltre, l'investimento per pulire, organizzare e rendere fruibili i dati ai modelli di ML garantisce un ritorno notevole. Dati curati aiutano i modelli ad apprendere in modo efficace generando previsioni accurate, risultati attendibili e, in sostanza, il successo dell'intero progetto di ML.

Tuttavia, il formato dei dati di addestramento dipende in larga misura dalla struttura di ML sottostante. I sistemi di ML tradizionali sono basati su disco e molti flussi di lavoro dei data scientist si concentrano sulla stesura di buone prassi e procedure di codifica manuale per gestire grandi quantità di file. Recentemente, i sistemi di ML basati su lakehouse usano uno store di funzionalità simile a un database e il workflow dei data scientisti oggi usa SQL come linguaggio primario. Di conseguenza, le strutture di dati in tabelle di qualità e ben costruite sono il formato di dati più fruibili e conveniente per i sistemi di ML.

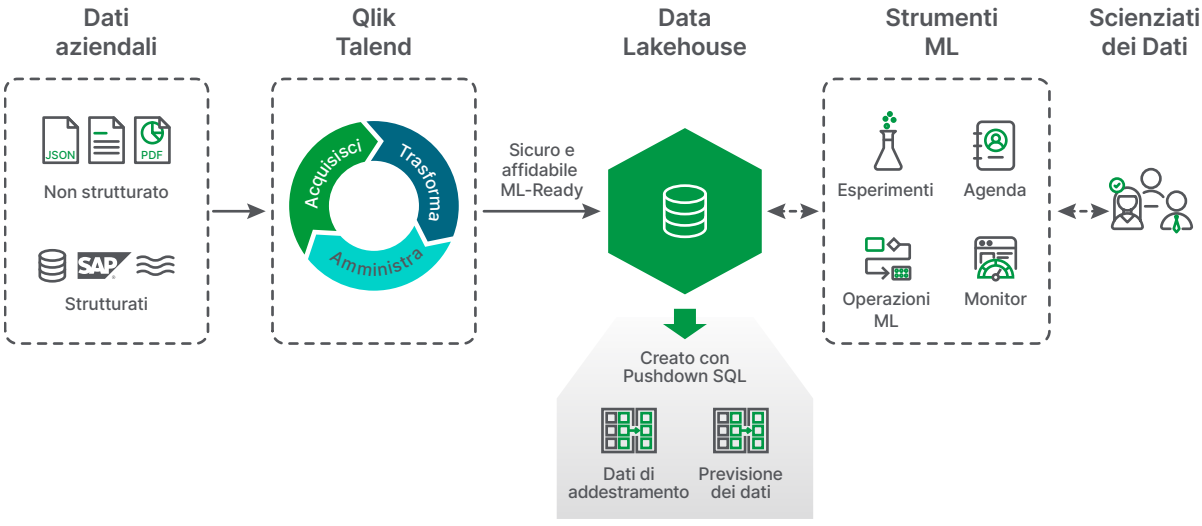


Figura 7. Rendere i dati fruibili per il ML

Rendere i dati fruibili dall'IA generativa

I modelli linguistici di grandi dimensioni (LLM) come GPT-4 di OpenAI, Claude di Anthropic, LaMDA di Google AI e Gemini sono stati pre-addestrati con grandi quantità di dati testuali e sono il cuore della GenAI. Si stima che il modello GPT-3 di OpenAI sia stato addestrato con circa 45 TB di dati, oltre 300 miliardi di token. Nonostante la quantità di input, i LLM non sanno rispondere a domande specifiche su un'azienda perché non hanno accesso ai dati specifici. La soluzione è implementare questi modelli con le proprie informazioni per ottenere applicazioni di IA più corrette, rilevanti e affidabili.

Il metodo per integrare i dati aziendali in un'applicazione basata su LLM è detto generazione potenziata dal recupero (RAG, Retrieval-Augmented Generation). Questa tecnica usa solitamente informazioni testuali ricavate da fonti non strutturate e basate su file come presentazioni, archivi di email, documenti di testo, PDF, trascrizioni ecc. I testi vengono suddivisi in blocchi facilmente gestibili e convertiti in rappresentazioni numeriche utilizzate dai LLM, attraverso un processo detto embedding. Queste rappresentazioni vengono poi archiviate in un database vettoriale come Chroma, Pinecone o Weviate. È interessante notare che molti fornitori di database tradizionali, come PostgreSQL, Redis, e SingleStoreDB, supportano anche quelli vettoriali. Inoltre, di recente anche le piattaforme in cloud come Databricks, Snowflake e Google BigQuer hanno aggiunto il supporto per i vettoriali.

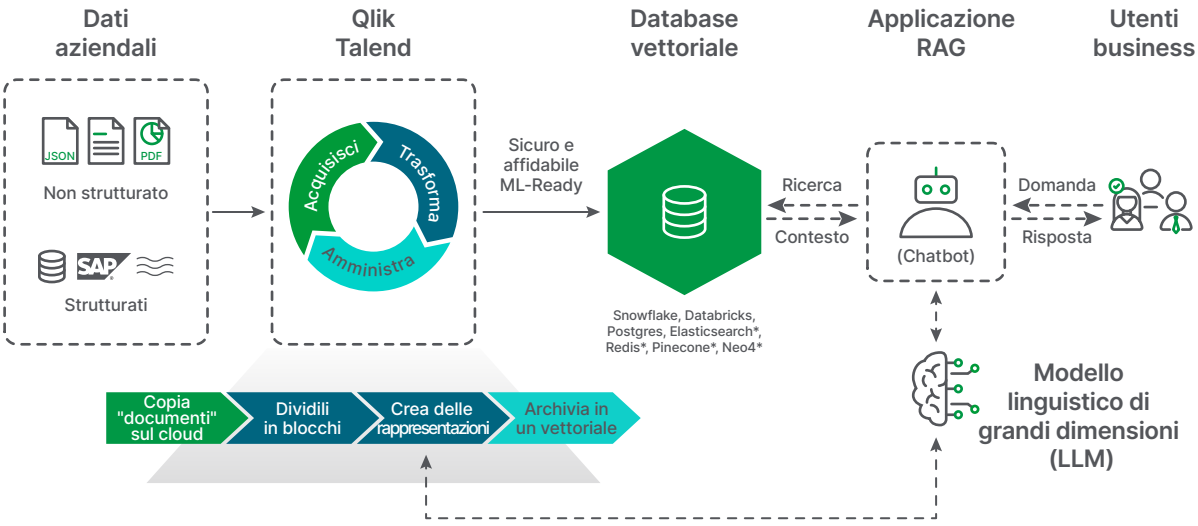


Figura 8. Rendere i dati fruibili dall'IA generativa

Indipendentemente dal fatto che i dati source siano strutturati o meno, l'approccio di Qlik garantisce una qualità dei dati immediatamente fruibile da applicazioni basate su LLM, GenAI o RAG.

AI Trust Score

Una volta definiti i sei criteri cardine per aver dati pronti e conformi, la domanda è: questi criteri si possono codificare e tradurre facilmente nella pratica quotidiana? E come si riconoscono a colpo d'occhio i dati pronti per l'IA? Una soluzione è quella di usare AI Trust Score di Qlik come indicatore complessivo della prontezza dei dati.

AI Trust Score assegna un parametro diverso a ogni criterio e poi aggrega questi valori per creare un punteggio complessivo: un modo rapido e affidabile per valutare se i dati sono pronti per l'IA. Inoltre, poiché i dati aziendali cambiano continuamente, il punteggio viene regolarmente controllato e ricalibrato per valutare le oscillazioni nei dati.

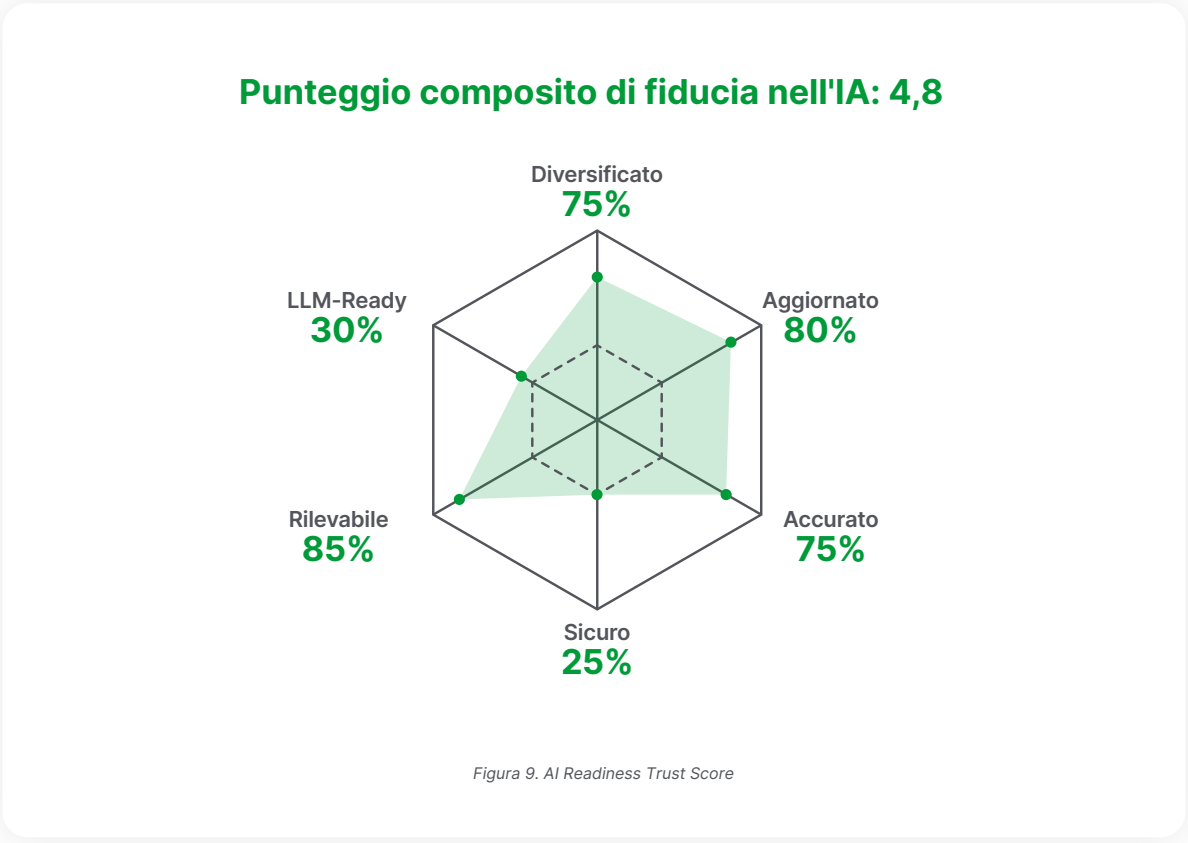


Figura 9. AI Readiness Trust Score

AI Trust Score combina varie metriche per creare un unico punteggio di prontezza dei dati, facile da comprendere.

La base di dati di Qlik Talend per l'IA

La necessità di avere dati di alta qualità e in tempo reale per prendere decisioni ponderate, aumentare l'efficienza operativa e sostenere l'innovazione non è mai stata così pressante. Per questo le aziende di successo scelgono le soluzioni di integrazione e qualità dei dati leader di mercato offerte da Qlik Talend, che forniscono dati affidabili a warehouse, lake e altre piattaforme in modo efficiente. La nostra offerta completa e di alto livello usa pipeline automatizzate, trasformazioni intelligenti e set di dati affidabili e di qualità per garantire l'agilità necessaria ai professionisti dei dati, unita alla governance e alla conformità richieste dalle aziende.

Se devi creare un warehouse o un lake per ottenere analisi accurate, modernizzare le strutture di dati operativi per aumentare l'efficienza in azienda o usare dati in multi cloud per progetti di intelligenza artificiale, Qlik Talend saprà indicarti la via.

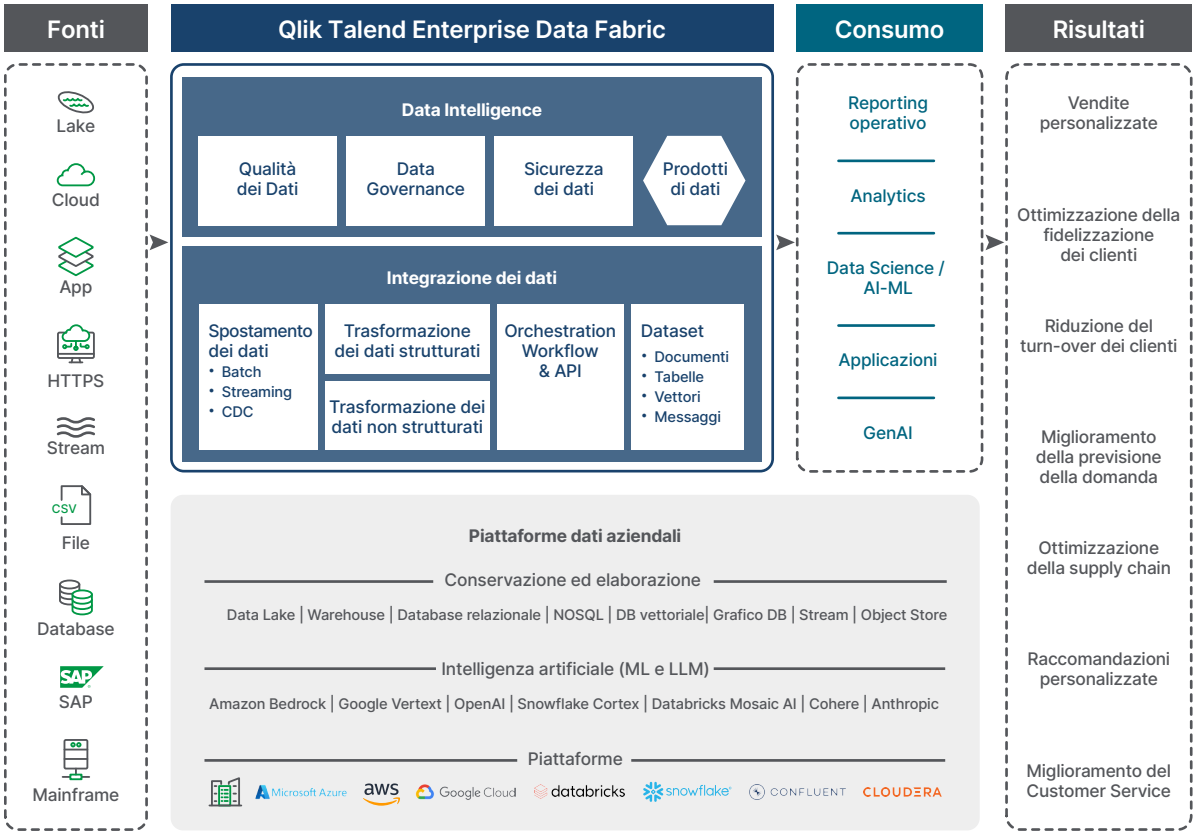


Figura 10. Qlik Talend Enterprise Data Fabric per IA e Analytics

Conclusione

Nonostante il potere rivoluzionario del machine learning e il potenziale di crescita dell'IA generativa, avere dati pronti per l'IA resta un elemento indispensabile per qualsiasi progetto di IA di successo. Questo white paper ha presentato i sei criteri fondamentali per creare una base di dati solida e affidabile che consenta alla tua azienda di sfruttare appieno il potenziale dell'IA.

Vuoi sfruttare appieno il potenziale dell'IA nella tua azienda?

Inizia da qui

Informazioni su Qlik

Qlik converte ambienti di dati complessi in informazioni fruibili, favorendo insight strategici per il business. Con oltre 40.000 clienti nel mondo, il suo portafoglio si avvale di AI/ML avanzati di livello enterprise che permettono di migliorare la data quality. Leader nell'integrazione e nella governance dei dati, Qlik offre soluzioni complete che funzionano con fonti di dati differenti. Le analisi intuitive e in tempo reale di Qlik scoprono modelli nascosti, consentendo ai team di affrontare sfide complesse e cogliere nuove opportunità. Gli strumenti IA/ML, pratici e scalabili, consentono di prendere decisioni migliori e più tempestive. Grazie ad una tecnologia avanzata e forti competenze del settore, indipendenti dalla piattaforma, Qlik è un partner strategico che rende i propri clienti più competitivi.

qlik.com

24-DQ-0001-01-WHM