Informatica

White paper

Intelligenza artificiale per l'azienda intelligente basata sui dati

Innovazioni basate su machine learning in CLAIRE apportano nuovi progressi nella gestione dei dati

Informazioni su Informatica

La Digital Transformation cambia le nostre aspettative: migliori servizi, consegne più rapide, il tutto a costi più contenuti. Le aziende devono trasformarsi per rimanere competitive e i dati sono la risposta per riuscirci.

Quale leader mondiale nell'Enterprise Cloud Data Management, possiamo supportarti per evolvere in modo intelligente in qualsiasi settore, categoria o nicchia di mercato. Informatica ti offre la possibilità di diventare più agile, realizzare nuove opportunità di crescita o persino inventare cose nuove. Siamo focalizzati al 100% sui dati e questo ti offrirà la flessibilità necessaria per competere ed avere successo.

Ti invitiamo a scoprire tutto quello che Informatica ha da offrirti, sprigionando "the power of data" per promuovere la tua prossima intelligent disruption.

Indice

"I leader di dati e degli analytics affrontano la complessità nel loro panorama di dati. Le previsioni delle nostre soluzioni per la gestione dei dati riconoscono gli sviluppi importanti e la crescente domanda di funzionalità del Cloud, di architetture dati connesse, dei metadati e dell'automazione di attività di routine e non di routine attraverso l'applicazione dell'intelligenza artificiale".[1]

— Gartner

L'importanza dell'intelligenza artificiale

Intelligenza artificiale (AI) e machine learning (ML) stanno alimentando la digital transformation in ogni settore in tutto il mondo. Secondo gli executive l'intelligenza artificiale è al primo posto tra le strategie di trasformazione delle loro attività. E ormai ha anche assunto un ruolo pervasivo nella vita di ogni giorno, dai film che guardiamo alle auto che guidiamo. Intelligenza artificiale e ML sono fondamentali per scoprire nuove terapie nella ricerca medica, ridurre le frodi e i rischi nei servizi finanziari e fornire customer experience davvero personalizzate.

Ai business leader, la combinata AI/ML può sembrare una specie di magia: anche se il suo potenziale impatto è chiaro, potrebbero non capirlo del tutto o non capire quale sia il modo migliore di esercitare tutta la sua potenza innovativa. Intelligenza artificiale e ML sono le tecnologie alla base di molte nuove soluzioni di business dedicate ad attività quali migliori prossime mosse, monitoraggio della customer satisfaction, efficienza operativa e innovazione dei prodotti. La tecnologia di machine learning soprattutto, e in particolare il deep learning, ha fame di dati. Per ottenere la precisione richiesta, ML ha bisogno di grandi quantità di dati per l'addestramento. Questi dati devono riflettere accuratamente lo stato attuale del business. L'intelligenza artificiale addestrata con dati errati o limitati avrà un impatto terribile sulle iniziative di business, al punto da invertire il risultato desiderato.

Per avere un'AI efficace, una in cui vengono utilizzate e addestrate le giuste funzionalità, è necessario attingere a un'ampia varietà di dati dall'interno e dall'esterno dell'organizzazione. Questi dati devono essere riuniti in modo da poter creare e addestrare un modello di ML. E questo richiede la gestione dei dati. Non si tratta solo di affrontare scala e complessità, ma anche fiducia. I dati utilizzati per addestrare il modello provengono dai sistemi giusti? Abbiamo rimosso le informazioni di identificazione personale (PII) e abbiamo rispettato tutte le normative? Agiamo con trasparenza e possiamo dimostrare il lineage dei dati utilizzati dal modello? Possiamo documentare ed essere pronti a mostrare alle autorità regolatorie o agli investigatori che non ci sono distorsioni nei dati? Tutto ciò richiede un buon controllo e una base di gestione dei dati. Senza una solida base di gestione dei dati, l'intelligenza artificiale risulta incomprensibile e inaffidabile; in altre parole, senza la gestione dei dati, l'intelligenza artificiale diventa una scatola nera dalle conseguenze indesiderate.

L'intelligenza artificiale ha bisogno di dati

Il successo dell'AI dipende dall'efficacia dei modelli progettati dai data scientist per addestrarla e scalarla. E il successo di questi modelli dipende dalla disponibilità di dati affidabili e tempestivi.

Perché i data scientist incaricati di creare modelli AI/ML hanno bisogno di dati di alta qualità? Prendiamo, ad esempio, un modello di previsione incaricato di anticipare il comportamento di un consumatore. Una caratteristica preziosa per un tale modello potrebbe essere la città del consumatore indicata dal codice postale. Cosa succede se i dati del codice postale sono mancanti, incompleti o imprecisi? Il comportamento del modello sarà influenzato negativamente sia durante l'addestramento sia durante l'implementazione, con il rischio di condurre a previsioni errate e ridurre il valore dell'intero sforzo. Inoltre, un codice postale accurato, completo e verificato potrebbe aiutare a prevedere la segmentazione del mercato, la classe di reddito, l'età, l'aspettativa di vita di un individuo e via dicendo, motivo in più per averne cura. Dovremmo tutti aspettarci che "l'intelligenza artificiale spiegabile" diventi un mandato regolamentato, non solo un'opzione. Senza il lineage e la tracciabilità basati sui metadati, le applicazioni e gli insight basati sull'intelligenza artificiale non possono essere implementati in produzione.

[1] Gartner, Previsioni 2020: Data Management Solutions, Rick Greenwald, Donald Feinberg, Mark Beyer, Adam Ronthal, Melody Chien, 5 dicembre 2019.

L'intelligenza artificiale ha bisogno di una gestione intelligente dei dati se si vogliono trovare rapidamente tutte le funzionalità del modello, trasformare automaticamente i dati per soddisfare le esigenze del modello AI (ridimensionamento delle funzionalità, standardizzazione, ecc.), deduplicare i dati e fornire dati master affidabili su clienti, pazienti, partner e prodotti, e fornire il lineage end-to-end dei dati, anche all'interno del modello e delle sue attività. Il successo dell'AI dipende dall'efficacia dei modelli progettati dai data scientist per addestrarla e scalarla. E il successo di questi modelli dipende dalla disponibilità di dati affidabili e tempestivi.

I dati hanno bisogno di intelligenza artificiale

AI e ML svolgono anche un ruolo fondamentale nel ridimensionare le pratiche di gestione dei dati. A causa degli enormi volumi di dati necessari per la digital transformation, le organizzazioni devono scoprire e catalogare i propri dati e metadati più rilevanti per certificarne la pertinenza, il valore e la sicurezza e per garantire la trasparenza. Devono pulire e padroneggiare questi dati. E devono governare e proteggere efficacemente questi dati. Se i dati non vengono gestiti in modo efficace e scalabile, i modelli AI/ML subiranno la stessa sorte di ogni iniziativa di data warehousing tradizionale degli ultimi 30 anni: partire da dati di scarsa qualità e rimanere con informazioni non affidabili.

Secondo una recente ricerca, il volume complessivo del traffico dei data center raggiungerà i 20,6 zettabyte nel 2021, e il numero di dispositivi e connessioni connessi supererà i 25 miliardi entro il 2022.[2] Tutti questi dati devono essere elaborati e resi utilizzabili e affidabili nel rispetto delle politiche di governance. A tutto ciò si aggiunge l'esigenza di muoversi rapidamente e rispondere ai cambiamenti nella strategia e nei processi di business. Lo sforzo necessario per preparare i dati per le iniziative di digital transformation è aumentato di complessità, di pari passo con la crescita dei dati. Secondo LinkedIn, la posizione di data scientist è uno dei lavori più promettenti negli Stati Uniti.[3] E il numero di data engineer ricercati dalle aziende ha recentemente registrato un aumento del 96% anno su anno.[4] Ma il solo impiego di questi ruoli non basterà a gestire l'aumento del volume dei dati.

Non adottare un approccio lineare a una sfida esponenziale

Non possiamo risolvere queste sfide semplicemente mettendo più ingegneri e sviluppatori a lavorare sul problema: perché non è un problema risolvibile su scala lineare e umana. Gli approcci tradizionali sono pieni di inefficienze. I progetti vengono implementati in silos con poca visibilità dei metadati end-to-end e automazione limitata. Non c'è apprendimento, l'elaborazione è costosa e le fasi di governance e privacy vengono ripetute più e più volte. Quindi, cosa dovrebbero fare le organizzazioni che vogliono muoversi alla velocità del business, abilitare il self-service, servire meglio i propri clienti, aumentare l'efficienza operativa e innovare rapidamente?

[2] Cisco, Global Cloud Index Forecast and Complete Visual Networking Index Forecast
[3] LinkedIn, "LinkedIn's Most Promising Jobs of 2019."
[4] Datanmi, "Data Engineering Continues to Move the Employment Needle."

È qui che l'intelligenza artificiale dà il meglio. L'intelligenza artificiale può automatizzare e semplificare le attività relative alla gestione dei dati, attraverso discovery, integrazione, pulizia, governance e mastering dei dati. I metodi di machine learning possono apprendere e assumere compiti banali e ripetitivi, lasciando liberi sviluppatori e utenti di lavorare su progetti innovativi e di alto valore. L'intelligenza artificiale migliora la comprensione dei dati e identifica la privacy dei dati e le anomalie nella qualità. L'intelligenza artificiale è un partner perfetto per sviluppatori, analisti, amministratori e utenti business, perché possono velocizzare le attività attraverso l'automazione e il potenziamento con consigli e migliori azioni successive.

L'intelligenza artificiale diventa più efficace se pensiamo a come possa aiutarci ad accelerare i processi end-to-end nell'intero ambiente di dati. Ecco perché consideriamo l'intelligenza artificiale essenziale per la gestione dei dati e perché Informatica® ha concentrato gli investimenti così pesantemente sull'engine CLAIRE®, la nostra funzionalità di intelligenza artificiale basata sui metadati. CLAIRE sfrutta tutti i metadati unificati di livello enterprise per automatizzare e scalare le attività di stewardship e gestione dei dati di routine.

Quattro principali vantaggi dell'intelligenza artificiale per la gestione dei dati
In generale, l'intelligenza artificiale avvantaggia i team di gestione dei dati in quattro modi: migliora la produttività dei professionisti dei dati, migliora l'efficienza delle attività, fornisce un'esperienza dei dati guidata in modo più intelligente e una comprensione più profonda e accelera i processi di data governance. Di seguito riportiamo alcuni esempi per mostrare ciò che è possibile ottenere oggi.

Produttività: Un sistema di raccomandazione per la data integration aiuta i data engineer a creare rapidamente mappature per estrarre, trasformare e fornire dati. Il suggeritore apprende dalle mappature esistenti, comprende il contenuto di business del database e del file system e suggerisce trasformazioni appropriate per la standardizzazione e la pulizia dei dati prima di consegnarli ai sistemi di destinazione e ai consumatori di dati.

Efficienza: In una tipica azienda vengono eseguiti ogni giorno migliaia di processi di data integration. Il monitoraggio di questi processi è in gran parte passivo, con strumenti di amministrazione che registrano solo il tempo impiegato e la CPU e la memoria consumate. L'intelligenza artificiale può apprendere dai valori storici di serie temporali nei file di log e di monitoraggio, e contrassegnare in modo proattivo i valori anomali, nonché prevedere i problemi che potrebbero verificarsi se non gestiti in anticipo.

Esperienza sui dati: Quando un'entità del mondo reale (ad esempio, una cartella clinica o un ordine di vendita) viene archiviata in un database o in un insieme di file, i suoi dati vengono distrutti e distribuiti in più tabelle o file, ottimizzati per l'archiviazione e le performance. L'intelligenza artificiale può rilevare le relazioni tra i dati e ricostituire rapidamente l'entità originale. Gli utenti non devono ricordare o cercare documentazione obsoleta sulle relazioni chiave primaria-chiave esterna e unire manualmente i vari set di dati. Inoltre, l'intelligenza artificiale può identificare set di dati simili e formulare raccomandazioni basate su modelli di utilizzo, qualità dei dati e collaborazioni da crowd-sourcing.

Data governance: Un passaggio comune ma noioso nella data governance consiste nell'associare termini di business a elementi di dati fisici per stabilire il contesto di business e la rilevanza per gli elementi di dati e rendere i dati comprensibili agli utenti. In molti casi, l'intelligenza artificiale può collegare automaticamente i termini di business ai dati fisici utilizzando una combinazione di tecniche di elaborazione del linguaggio naturale (NLP, Natural Language Processing) e identificazione del tipo di business. Ciò può ridurre drasticamente la fatica di questo compito soggetto a errori. Nell'attuale era del Cloud, è importante sottolineare come questo approccio sia valido anche per le applicazioni SaaS. I metadati possono essere raccolti da applicazioni SaaS come ad esempio Salesforce e Workday e successivamente aggiunti al catalogo di livello enterprise.

Gestione dei dati basata sull'intelligenza artificiale: un esempio dal settore bancario
Per illustrare perché l'intelligenza artificiale ha bisogno della gestione dei dati e perché i dati hanno bisogno dell'intelligenza artificiale, consideriamo un esempio bancario.

Applicando l'intelligenza artificiale a un numero sempre maggiore di dati per analytics avanzati, predittivi e in tempo reale, le banche possono:
- Offrire servizi più personalizzati che aumentano la fidelizzazione dei clienti
- Ridurre le transazioni fraudolente al point-of-sale
- Aumentare i risultati degli investitori riducendo i costi dei consulenti patrimoniali
- Ridurre il costo della conformità alle normative per progetto

Dal punto di vista della gestione dei dati, l'intelligenza artificiale può rilevare e catalogare automaticamente tutti i tipi di dati rilevanti come ad esempio ERP, CRM, app Cloud e Web, file di log e macchine, dati di terze parti e così via. Questo comporta per i data scientist un vantaggio nell'accedere a tutti i dati di cui hanno bisogno per eseguire centinaia di esperimenti alla ricerca di modelli che rivelino informazioni sul comportamento dei consumatori, su attività fraudolente, su opportunità di investimento abbinate alla propensione al rischio dei consumatori e altro ancora.

L'intelligenza artificiale, in relazione alla gestione dei dati, può arricchire automaticamente una vista a 360 gradi dei clienti e delle persone di interesse (POI, Person of Interest) scoprendo le relazioni tra i dati dei clienti e abbinando le informazioni a persone specifiche. In questo modo le organizzazioni possono interagire meglio con i propri clienti con offerte più pertinenti ed esperienze senza interruzioni su canali di vario tipo, online, mobili o via telefono. Una vista a 360 gradi delle POI aiuta le banche a scoprire schemi e reti di attività fraudolente molto più velocemente, risparmiando, potenzialmente, milioni.

E l'intelligenza artificiale può automatizzare e guidare le attività di data integration e data quality per combinare e ripulire i dati da centinaia di fonti, aumentando così il potere predittivo di modelli e algoritmi analitici. È stato dimostrato che dati più numerosi e migliori, combinati con AI/ML e analytics avanzati, producono risultati significativi, come ad esempio il miglioramento delle offerte e l'identificazione delle frodi.

L'intelligenza artificiale potenzia anche la data governance che garantisce che le policy non siano solo documentate ma effettivamente applicate. Ciò aiuta i professionisti della sicurezza delle informazioni a rispettare le normative sulla privacy dei dati come ad esempio il Regolamento generale sulla protezione dei dati (GDPR), il Sarbanes-Oxley Act (SOX), Basilea II, Basilea III e altro ancora.

Informatica CLAIRE: la parte "intelligente" in Intelligent Data Management Cloud

Ecco l'approccio di Informatica alla promozione della produttività del data management con il machine learning.

1. Intelligent Data Management Cloud™: abbiamo creato una piattaforma integrata per la gestione dei dati end-to-end che garantisce la massima produttività. Offrendo connettività unificata, metadati e gestione delle attività, la piattaforma unificata accelera lo sviluppo e l'implementazione dei nuovi progetti di gestione dei dati. La piattaforma offre una serie avanzata e coerente di funzionalità per la gestione dei dati tra fonti dati on-premise, Cloud, multi-Cloud e multi-ibride. Chiamiamo questa piattaforma unificata di gestione dei dati Intelligent Data Management Cloud.

La piattaforma è modulare: puoi partire con qualsiasi tool individuale e crescere secondo il tuo ritmo.

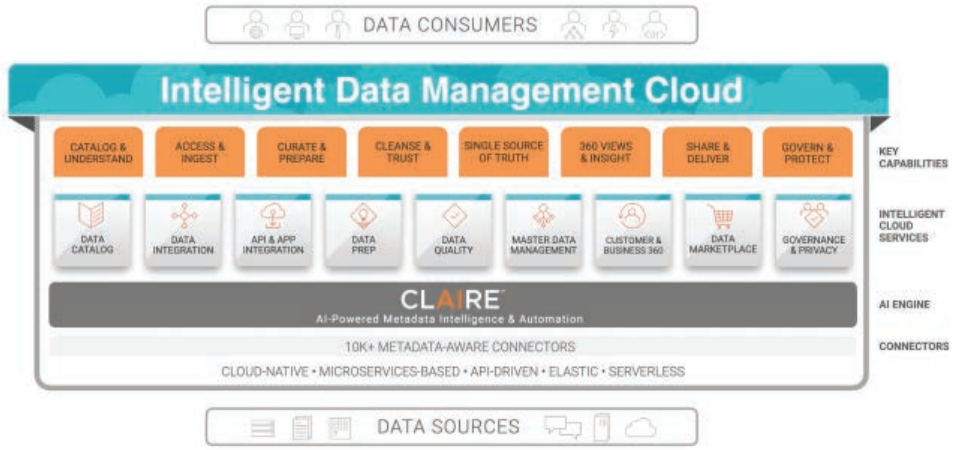

Figura 1. Intelligent Data Management Cloud integra funzionalità di gestione dei dati con connettività condivisa, conoscenze operative, data intelligence e metadata intelligence.

2. Metadati: Informatica da tempo è nota quale leader per la gestione dei metadati tecnici e di business. Informatica ha oggi ampliato le sue funzionalità in questo settore raccogliendo uno spettro più ampio di metadati di tutta l'azienda, compresi:
   - Metadati tecnici, come ad esempio tabelle di database, informazioni sulle colonne, statistiche del profilo dei dati, script e lineage dei dati
   - Metadati di business, che acquisiscono il contesto dei dati, il relativo significato, la rilevanza e l'importanza per i diversi processi e funzioni di business
   - Metadati operativi sui sistemi e sull'esecuzione dei processi per rispondere a domande quali: quando sono stati aggiornati i dati l'ultima volta, o il momento in cui è stato eseguito per l'ultima volta il processo di caricamento oppure ancora i dati ai quali si accede di più
   - Metadati sull'utilizzo riguardanti l'attività degli utenti, inclusi i set di dati ai quali è stato effettuato l'accesso, i risultati delle ricerche selezionati e le classificazioni o i commenti forniti

Questa più ampia raccolta di metadati è fondamentale per il machine learning. Fornisce set di dati che vengono utilizzati per "addestrare" gli algoritmi di machine learning e consente loro di adeguarsi e produrre risultati migliori.

3. Intelligenza: Informatica offre una combinazione integrata di metadati e AI/machine learning con CLAIRE.

   I metadati raccolti da Intelligent Data Management Cloud offrono un ampio bagaglio di informazioni che gli algoritmi di CLAIRE possono utilizzare per conoscere il panorama dei dati di un'azienda. Queste conoscenze aiutano CLAIRE a offrire consigli intelligenti, automatizzare lo sviluppo e il monitoraggio dei progetti di gestione dei dati e adattarsi ai cambiamenti sia dentro sia fuori l'azienda. CLAIRE è ciò che guida l'intelligenza di tutte le funzionalità di gestione dei dati all'interno di Intelligent Data Management Cloud.

CLAIRE aiuta una vasta gamma di utenti:

- I data engineer troveranno molte attività di implementazione in parte o completamente automatizzate
- I data analyst potranno individuare e preparare in modo più semplice i dati a loro necessari
- Gli utenti business identificheranno più rapidamente i dati che devono essere sottoposti ai controlli di conformità e data governance stabiliti
- I data scientist comprenderanno dati in modo più rapido
- I data steward capiranno che è più semplice visualizzare la data quality
- I professionisti della sicurezza e della privacy dei dati troveranno più semplice rilevare l'uso improprio dei dati, proteggere i dati sensibili e dimostrare che i dati sono adeguatamente controllati
- Gli amministratori e gli operatori potranno contare sulla manutenzione predittiva e l'ottimizzazione delle performance dei processi di gestione dei dati

Ecco alcuni esempi di come viene utilizzata oggi l'intelligenza di CLAIRE.

CLAIRE per la catalogazione dei dati

Scoprire e comprendere i dati in tuo possesso è il primo passo di qualsiasi iniziativa fondata sui dati. CLAIRE fornisce un engine di discovery basato su machine learning per scansionare e catalogare le risorse di dati in tutta l'azienda. Un intelligent data catalog basato su CLAIRE può aiutare data scientist, analisti e data engineer a trovare e consigliare i dati di cui hanno bisogno, riducendo significativamente il tempo impiegato nella discovery e nella preparazione dei dati.

Discovery avanzata delle relazioni

Un'attività chiave di catalogazione e modellazione dei dati è documentare le relazioni tra i set di dati. CLAIRE utilizza tecniche di machine learning per identificare automaticamente chiavi primarie, chiavi univoche e join tra set di dati strutturati. Così facendo riduce mesi di lavoro a pochi minuti. CLAIRE migliora continuamente la sua capacità di identificare le relazioni includendo gli esseri umani nel processo di cura dei dati. Ad esempio, gli utenti possono accettare o rifiutare le relazioni dedotte e CLAIRE impara da queste azioni.

Ad esempio, un analista di dati presso una banca che crea un report sui clienti con maggiori probabilità di rispondere a una campagna di marketing dovrebbe essere in grado di trovare prodotti esistenti e informazioni sui prestiti relative a tutti i clienti. Tuttavia, data la natura isolata dei dati all'interno dell'azienda, potrebbe fare fatica a trovare tali set di dati in più dipartimenti e store di dati. CLAIRE utilizza join documentati nei database, join eseguiti in altri strumenti come BI ed ETL e statistiche derivate dai valori dei dati per dedurre e consigliare i join all'analista dei dati. Questo aiuta ad espandere l'analisi dell'utente e utilizza tutte le informazioni disponibili per trovare l'audience di riferimento giusta per la campagna.

CLAIRE combina diverse tecniche per la discovery di chiavi e join. Statistiche di profilazione come unicità, conteggi null, metadati di colonna (ad esempio, nomi di colonna contenenti "ID") vengono combinate per scoprire chiavi primarie e univoche. I join e l'inferenza della chiave di join utilizzano quindi una combinazione di tecniche di machine learning come l'analisi della firma delle colonne per scoprire i join su larga scala in molti potenziali set di dati.

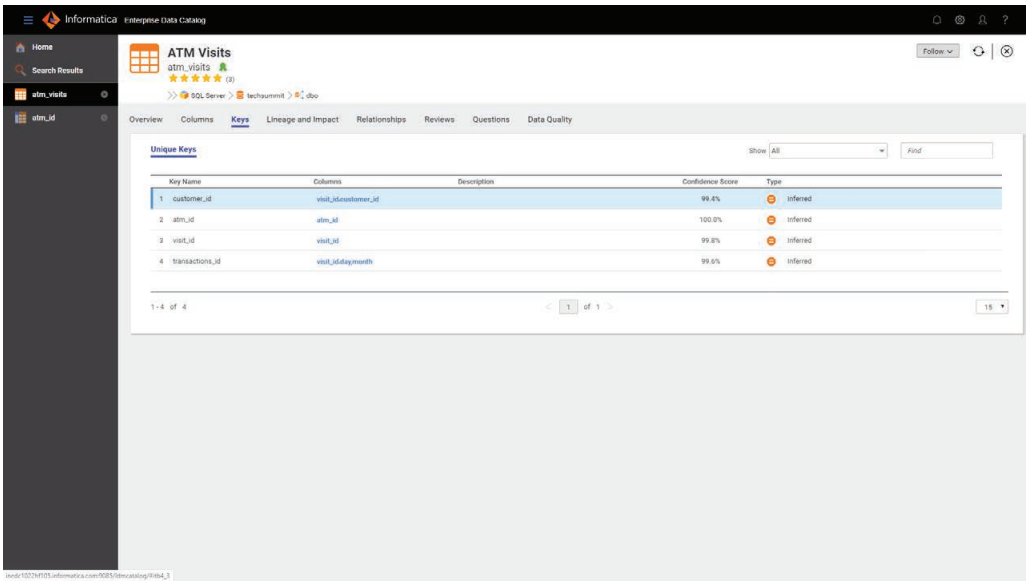

Figura 2. Scoprire chiavi univoche attraverso l'inferenza utilizzando tecniche di machine learning.

Identificazione intelligente della somiglianza tra dati

CLAIRE utilizza tecniche di machine learning come il clustering per rilevare dati simili tra migliaia di database e insiemi di file. L'identificazione intelligente della somiglianza tra dati è tra le funzionalità più importanti per tutta una serie di scopi, come ad esempio identificare i dati, rilevare duplicati, combinare singoli campi di dati in entità di business, propagare tag tra i set di dati e raccomandare set di dati agli utenti.

La somiglianza tra dati calcola quanto i dati di due colonne sono identici. Un approccio di tipo "brute force" per cercare di confrontare tutte le coppie di due colonne in un contesto enterprise (ad esempio in 100 milioni di colonne) sarebbe proibitivo dal punto di vista computazionale. Al contrario, la somiglianza dei dati utilizza tecniche di machine learning per raggruppare colonne simili e identificare le corrispondenze probabili.

Il processo si svolge in più fasi. In primo luogo, le colonne vengono raggruppate in cluster sulla base delle proprie caratteristiche. Successivamente, viene calcolata la sovrapposizione di dati in ciascun cluster. Infine, le coppie più promettenti vengono scelte per il calcolo della somiglianza dei dati utilizzando i coefficienti di Bray-Curtis e Jaccard.

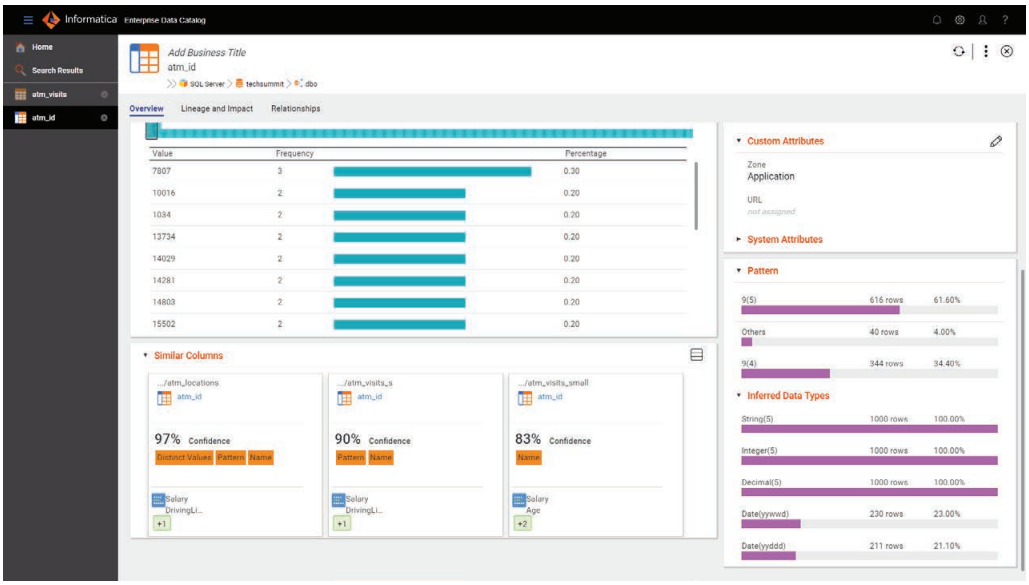

Figura 3. Identificazione di colonne simili utilizzando il clustering e i coefficienti di Bray-Curtis e Jaccard.

Discovery intelligente dei domini con i tag

CLAIRE è in grado di classificare i campi di dati applicando etichette semantiche a ciascuna colonna. Queste etichette semantiche sono dette domini di dati.

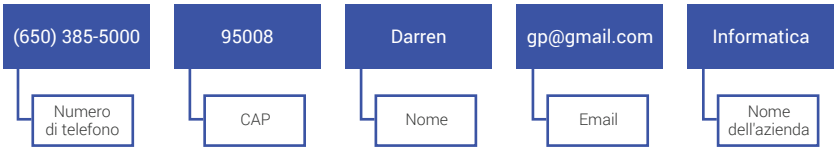

Figura 4. CLAIRE classifica automaticamente i campi dati e applica etichette semantiche chiamate tag.

Di norma le etichette semantiche si applicano valutando le regole sulla base di espressioni regolari, tabelle di riferimento o altre logiche complesse tramite codice scritto a mano. La definizione e la gestione di migliaia di regole simili è un processo tedioso.

CLAIRE utilizza invece il concetto dei tag per semplificare drasticamente il processo di discovery e di etichettatura dei campi di dati. Per le colonne non ancora classificate, l'utente non deve fare altro che fornire un semplice tag (ad esempio, "Data pagamento reclamo") che indica il contenuto della colonna. Il sistema apprende per associazione, propagando quindi automaticamente questo tag a tutte le colonne simili. Il "riconoscimento facciale" per i dati equivale i tag sui volti nelle foto di Facebook, con la differenza che volti identici vengono taggati in milioni di altre foto.

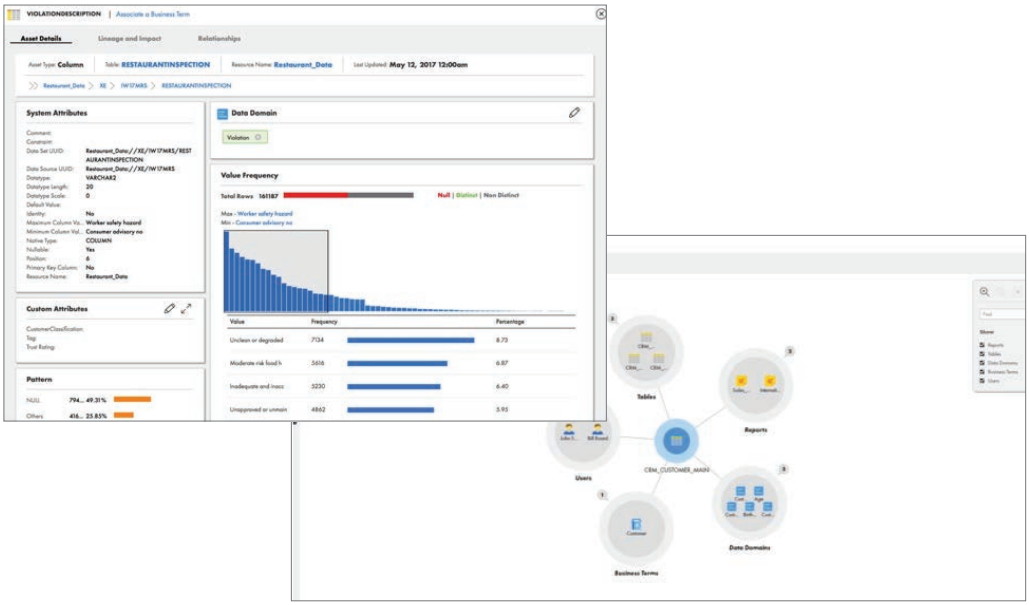

Figura 5. Classificazione automatica dei dati.

Discovery intelligente delle entità

Una volta identificati i domini delle colonne, CLAIRE può assemblare i singoli campi in entità di business di livello superiore. L'esempio che segue mostra un'entità relativa a un ordine di acquisto, creata combinando i campi identificati relativi al cliente e al prodotto. La discovery delle entità apprende dal modo in cui gli utenti hanno assemblato campi di dati disparati nei loro processi di analytics o data integration e applica questo apprendimento per derivare entità nel panorama dei dati di livello enterprise.

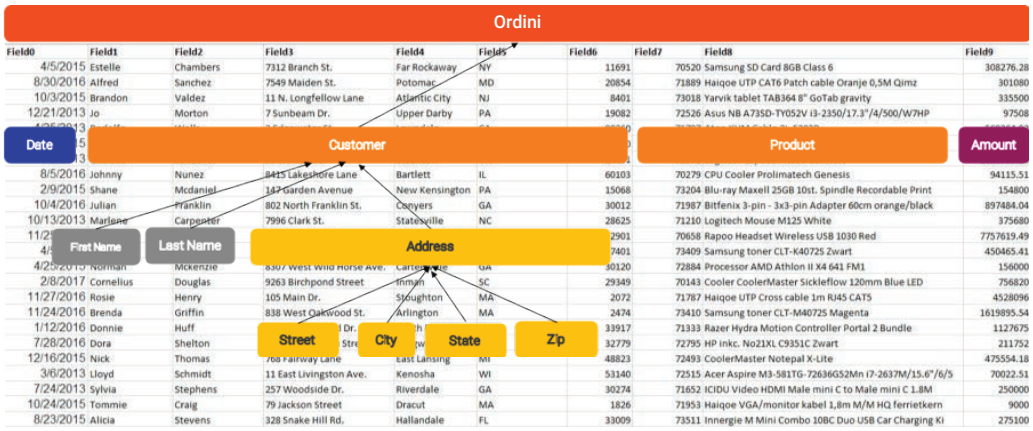

Figura 6. Combinazione dei domini di dati per rilevare le entità da tabelle e file.

CLAIRE per gli analytics

L'automazione e l'intelligenza basate su CLAIRE accelerano notevolmente gli approfondimenti e i processi analitici, aumentano la disponibilità dei dati e semplificano la preparazione dei dati per gli analytics. CLAIRE migliora la produttività del data engineering con suggerimenti sulla pipeline dei dati e la capacità di analizzare automaticamente dati complessi e multistrutturati.

Consigli per la trasformazione

Chiudi il ciclo di progettazione e migliora la produttività dei data engineer con la creazione automatizzata di mappe per la data integration contenenti previsioni di trasformazioni ed espressioni. Quando un'organizzazione sceglie di ricevere consigli basati su CLAIRE, vengono analizzati i metadati anonimi dalle pipeline di dati dell'organizzazione e viene applicata l'intelligenza artificiale/ML per offrire consigli di progettazione. Questi metadati vengono utilizzati per generare suggerimenti per la trasformazione e l'espressione. CLAIRE migliora a ogni utilizzo, sia con l'accettazione che con il rifiuto della raccomandazione. Ciò accelera lo sviluppo, automatizza le attività ripetitive e consente a più tipi di utenti di connettersi e integrare rapidamente i dati.

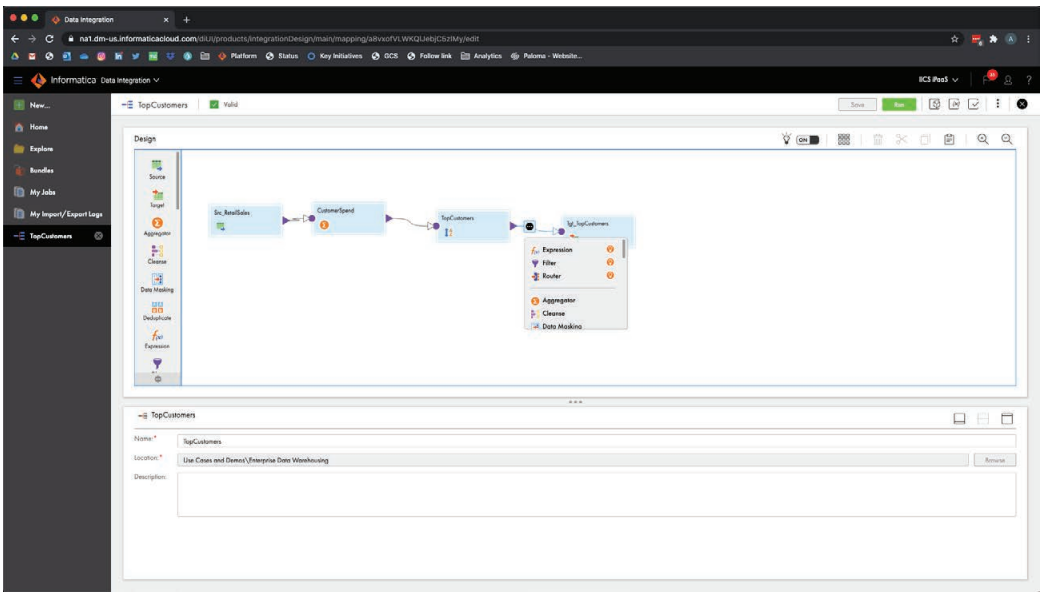

Figura 7. CLAIRE consiglia le migliori successive trasformazioni durante la creazione di pipeline di dati.

Esecuzione ottimizzata del processo su larga scala

CLAIRE utilizza una varietà di metodi di ottimizzazione per aumentare le performance di integrazione in tutta la pipeline dei dati. Un ottimizzatore intelligente decide il miglior engine di elaborazione per eseguire un carico di lavoro di big data in base alle caratteristiche delle performance; i suggerimenti di mappatura vengono presentati agli ingegneri dei dati in base alle passate attività degli utenti, mentre un ottimizzatore basato sui costi, insieme all'euristica, modifica in modo intelligente l'ordine di unione in una pipeline di dati per ottenere performance ottimali. Questi sono solo alcuni esempi di come CLAIRE ottimizza le pipeline di dati.

Raccomandazioni per colonne di join

CLAIRE suggerisce automaticamente le colonne di join (ovvero le chiavi di unione) quando un utente sceglie l'azione per combinare due set di dati. Ciò consente agli analisti di dati di risparmiare centinaia di ore di lavoro manuale nel tentativo di determinare il modo migliore di unire i set di dati in un insieme composito da analizzare. CLAIRE inizia con le relazioni di chiave primaria e chiave esterna (cioè, Pk-Fk) definite nei sistemi sorgente originali (database relazionali come ad esempio Oracle) dei set di dati che sono stati importati nel data lake. Se gli stessi set di dati vengono uniti in altri progetti, queste informazioni sulla colonna di unione verranno utilizzate anche per i suggerimenti. Tutte queste informazioni vengono elaborate e classificate da CLAIRE per suggerire le migliori colonne di join tra due set di dati. Inoltre, in base al campionamento dei set di dati, viene mostrata anche la percentuale di sovrapposizione dei dati tra le colonne suggerite.

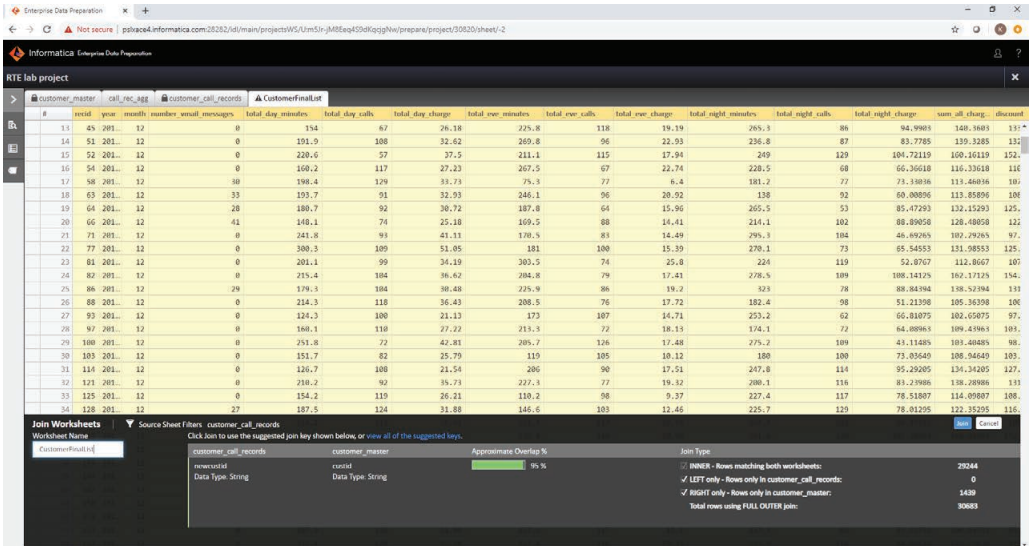

Figura 8. Suggerimenti automatici per le colonne di join quando si combinano due set di dati.

Raccomandazioni per la visualizzazione di Apache Zeppelin

Informatica Enterprise Data Preparation utilizza Apache Zeppelin per visualizzare i fogli di lavoro sotto forma di taccuino contenente grafici e diagrammi. Quando l'utente apre il taccuino di una pubblicazione, può leggere i consigli di visualizzazione di CLAIRE. Quando l'utente apre il taccuino per la prima volta dalla pubblicazione, vede le visualizzazioni dell'istogramma delle colonne numeriche derivate. Se la pubblicazione non contiene colonne numeriche derivate, l'utente vede una query di tabella "SELECT * FROM" nel primo paragrafo del taccuino.

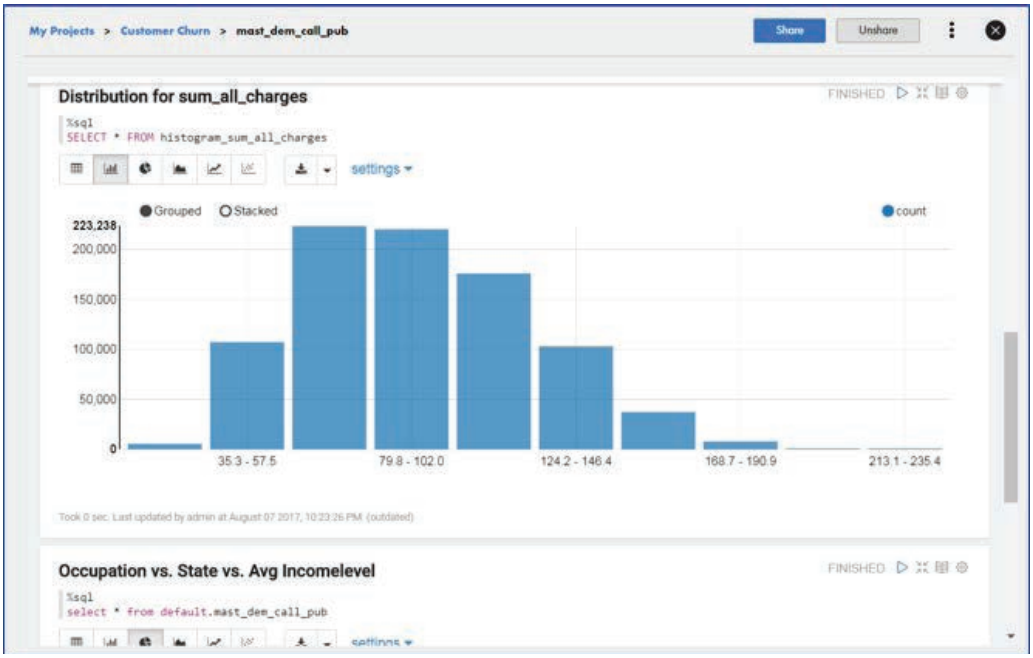

Figura 9. Visualizzazione consigliata nel taccuino Apache Zeppelin.

Consigli intelligenti sui dati

CLAIRE offre ai data analyst e data scientist i suggerimenti relativi a quali set di dati utilizzare nei propri progetti. Osserva i set di dati che gli utenti hanno selezionato e ne suggerisce di simili e meglio classificati, oppure propone altri set di dati che possano integrare quelli in uso. I consigli intelligenti sui dati aiutano gli utenti a evitare di ripetere lo stesso lavoro che molti altri colleghi potrebbero aver già svolto. I consigli comprendono:

- Una versione preparata degli stessi dati (dati sostituibili)
- Un'altra tabella che contiene gli stessi tipi di record (dati che è possibile sottoporre a unione)
- Una tabella che può essere sottoposta a join per essere arricchita con attributi di dati aggiuntivi (dati che è possibile sottoporre a join)

I consigli sui dati utilizzano tecniche di filtraggio basate sul contenuto per fornire suggerimenti su ulteriori set di dati. Le caratteristiche (termini) utilizzate per i set di dati comprendono informazioni su lineage, classificazione degli utenti e somiglianza dei dati. Numerose misure di somiglianza vengono utilizzate per assegnare il punteggio di equivalenza tra diversi set di dati. Tali punteggi vengono poi impiegati per consigliare set di dati con proprietà simili. Gli elementi complementari sono consigliati eseguendo query nel grafico dei metadati per identificare i set di dati comunemente utilizzati insieme da utenti diversi.

Intelligent Structure Discovery

Una quantità crescente di dati viene generata e raccolta tra macchine, aziende e applicazioni in formato non strutturato o non relazionale. Questi tipi di dati sono caratterizzati non solo dai grandi volumi, ma anche dalla loro velocità, varietà e variabilità. "Data drifting" è il termine oggi comunemente usato per descrivere la fluttuazione del formato, del ritmo e del contenuto dei dati in questi nuovi tipi di dati.

Informatica Intelligent Structure Discovery (ISD) con tecnologia CLAIRE è progettato per automatizzare l'acquisizione dei file e il processo di onboarding in modo che le aziende possano scoprire e analizzare file complessi. ISD fornisce supporto out-of-the-box per una varietà di formati di file di dati, tra cui clickstream, log IoT, CSV, testo delimitato, XML, JSON, Excel, ORC, Parquet, Avro, moduli PDF e file di tabelle di Word. CLAIRE può derivare automaticamente la struttura da questi file, rendendoli più facili da capire e da elaborare. Attraverso un approccio basato sul contenuto per il parsing dei file, può adattarsi alle modifiche frequenti dei file senza alcun impatto sull'elaborazione degli stessi.

ISD utilizza un algoritmo genetico per automatizzare il riconoscimento dei pattern all'interno dei file. Questo approccio utilizza il concetto di "evoluzione" per migliorare i risultati. Ogni soluzione candidata dispone di un insieme di proprietà che possono essere modificate e successivamente testate per stabilire se forniscono una soluzione più adatta. A tali strutture viene quindi assegnato un punteggio in funzione di diversi fattori, come ad esempio la copertura dell'input e i domini derivati. Le strutture con il punteggio maggiore entrano in una fase di "mutazione" dove vengono apportate diverse modifiche, ad esempio combinando sottostrutture per verificare se il punteggio migliora. Il processo si conclude quando viene stabilita l'adeguatezza della struttura rispetto ai dati.

ISD utilizza anche meccanismi personalizzati NER (Named Entity Recognition, riconoscimento delle entità denominate) e NLU (Natural Language Understanding, comprensione del linguaggio naturale) basati su ML per identificare i campi e i tipi di campo, il che semplifica le successive integrazioni e consente alle applicazioni esterne di utilizzare ISD come engine NER/NLU sottostante. Ad esempio, ISD viene utilizzato per rilevare le informazioni PII nel payload API in entrata e in uscita e facilita la conformità alle normative e una maggiore sicurezza aziendale. ISD viene utilizzato anche nei casi d'uso di discovery, importazione e streaming dei dati.

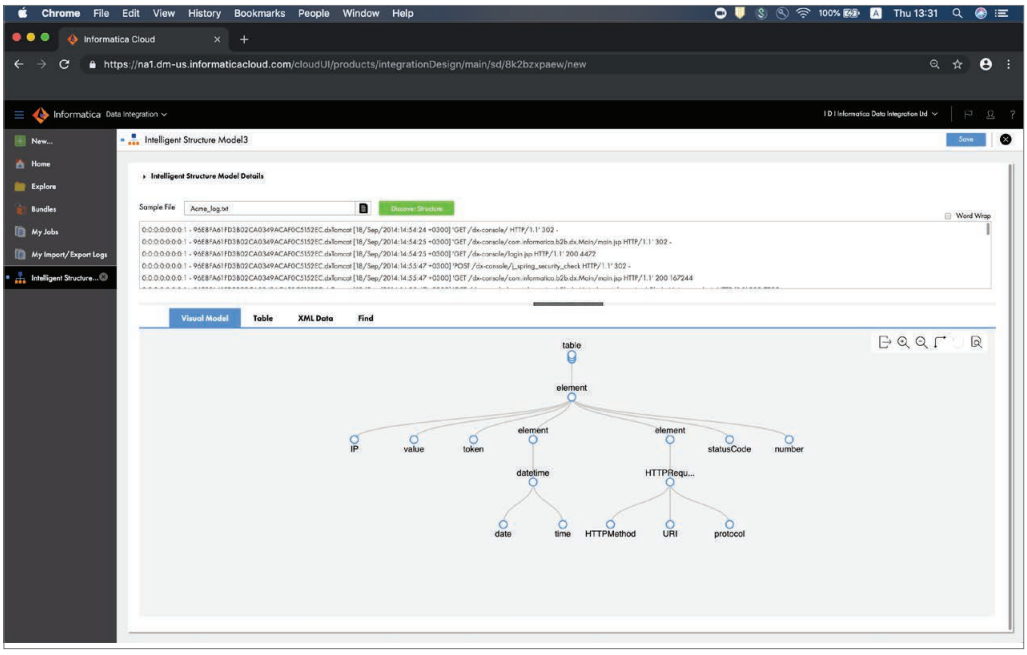

Figura 10. Rilevamento intelligente della struttura in file di dati non strutturati.

CLAIRE per il Master Data Management

L'automazione e l'intelligenza basate su CLAIRE con intelligenza artificiale avanzata e machine learning arricchiscono e migliorano l'accuratezza delle viste a 360 gradi su clienti, prodotti, fornitori e altri domini. Una varietà di tecniche miste AI/ML, che vanno dagli algoritmi deterministici, euristici e probabilistici alla corrispondenza di sintesi contestuale e alla corrispondenza delle entità di apprendimento attivo, vengono impiegate per fornire una corrispondenza di record rapida, scalabile e altamente accurata e l'arricchimento dei dati master.

Corrispondenza di sintesi

La sintesi è una tecnica di abbinamento di nuova generazione che affronta, ad esempio, la corrispondenza tra clienti e prospect, la corrispondenza tra interazioni/dati non strutturati e clienti, e scopre relazioni non ovvie. Utilizza "attributi contestuali", machine learning, NLP e una combinazione di corrispondenza probabilistica con regole dichiarative per raggiungere questo obiettivo.

Gli attributi demografici (ad esempio nome, indirizzo e numero di telefono), gli attributi di interazione (ad esempio email, chat Web) e gli attributi contestuali (ad esempio quando, cosa, dove, chi) sono potenti nell'abbinare i dati relativi al cliente con una data confidenza di livello. La NLP può sfruttare gli "attributi contestuali" dal testo non strutturato per fornire molti più punti dati da utilizzare nel processo di corrispondenza. Un algoritmo ML può essere molto efficace nella corrispondenza quando si utilizza un approccio di addestramento supervisionato in cui i data steward e gli esperti in materia etichettano un insieme opportunamente selezionato di coppie di corrispondenza come corrispondenza o non corrispondenza. Queste coppie di corrispondenza etichettate formano un set di addestramento utilizzato per produrre un algoritmo di corrispondenza.

Synthesis unirà una vista completa del cliente a 360 gradi composta da dati demografici, account, transazioni, interazioni e dati non strutturati. Gli algoritmi di corrispondenza tradizionali uniscono i record per formare un'unica vista del cliente, mentre la corrispondenza di sintesi gestisce tutti i dati del cliente in un grafico. I dati sono correlati insieme ai livelli di confidenza, dove è quindi possibile fornire più viste, o "prospettive", di un cliente.

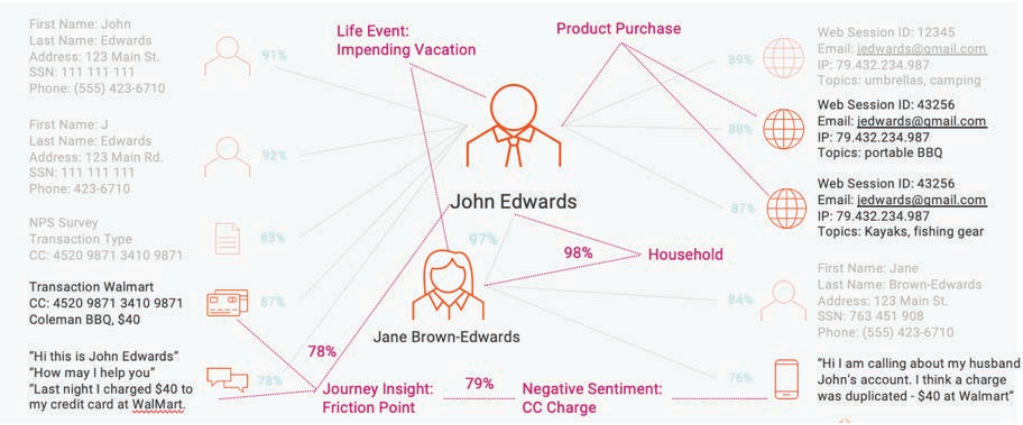

Figura 11. La corrispondenza di sintesi e il ragionamento deducono l'intelligenza che viene quindi archiviata come parte del Customer 360.

Corrispondenza delle identità
La corrispondenza dell'identità NAME3 di CLAIRE racchiude oltre 30 anni di addestramento e messa a punto utilizzando una varietà di tecniche come ad esempio la generazione di chiavi intelligenti per l'indicizzazione e il blocco, la stabilizzazione semantica del testo e confronto dei dati di parti e posizione, elenchi di modifica e regole di stabilizzazione del testo per 80 popolazioni e ponderazione intelligente dell'importanza delle caratteristiche per scopi diversi. Queste potenti tecniche consentono l'indicizzazione e il blocco su più campi, regole di corrispondenza e anti-corrispondenza definite dal cliente in base a requisiti e regole di corrispondenza e anti-corrispondenza definite dall'implementazione per integrare altre regole di intelligenza artificiale.

Corrispondenza delle entità
La corrispondenza delle entità trova i record di dati che fanno riferimento alla stessa entità del mondo reale (ad esempio clienti, prodotti e così via). I record di dati possono essere non strutturati (ad esempio informazioni sui clienti nascoste in una chat Web) e strutturati. La classificazione delle corrispondenze confronta una coppia di corrispondenze e determina se esiste una corrispondenza, una corrispondenza probabile o una mancata corrispondenza, insieme a un livello di confidenza. Esistono tecniche che utilizzano regole configurate manualmente (ovvero regole dichiarative) o regole AI (ovvero una configurazione appresa dalla macchina). I migliori risultati di corrispondenza si ottengono quando queste due tecniche vengono mescolate insieme.

Le regole dichiarative, create da esperti in materia, completano le potenti regole dell'AI che CLAIRE impiega sotto forma di classificatore informato di foreste casuali. CLAIRE utilizza l'apprendimento attivo supervisionato (al contrario dell'apprendimento in crowdsourcing o multiutente) per accelerare il processo di addestramento AI in cui vengono presentati a un utente microbatch di 10 o 20 coppie di corrispondenza per l'etichettatura (ad esempio, corrispondenza, corrispondenza probabile, mancata corrispondenza). Una volta etichettate, CLAIRE riqualifica il classificatore di foreste casuali e determina le migliori coppie di corrispondenza successive da presentare all'utente in questo processo di etichettatura iterativo. CLAIRE utilizza le coppie etichettate per dedurre le regole di blocco (ovvero rimuovere le non corrispondenze ovvie), eseguire il blocco, addestrare un modello ed eseguire la corrispondenza delle entità.

CLAIRE utilizza una combinazione di confronti/somiglianze di stringhe come ad esempio Jaccard, regole dichiarative derivate dalla profilazione dei dati, set di dati stabilizzati (file di popolazione, nickname, confronti semantici, ecc.) e regole definite dall'utente che gestiscono le eccezioni. Queste regole dichiarative affrontano le lacune e le eccezioni e aiutano ad accelerare il processo di formazione di apprendimento attivo (cioè a ridurre il numero di coppie di corrispondenza necessarie per l'apprendimento), ad accelerare la creazione di funzionalità delle regole di intelligenza artificiale e ad aumentare la precisione della corrispondenza. Ad esempio, ogni volta che nome, data di nascita e codice fiscale si riflettono fortemente, la regola li classifica come corrispondenza. Questa combinazione di regole dichiarative e regole di intelligenza artificiale accelera la formazione e migliora l'accuratezza della corrispondenza.

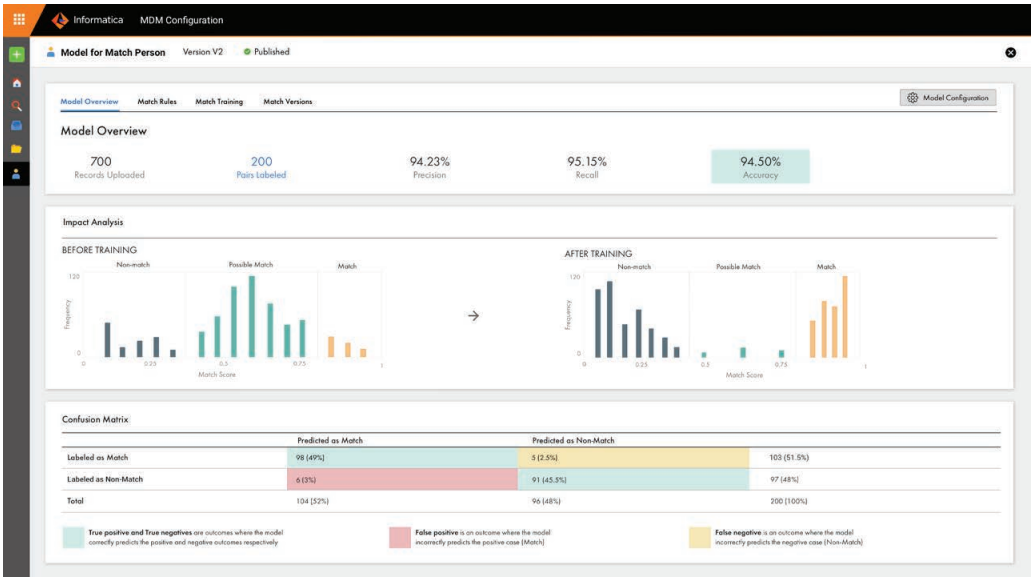

Figura 12. Corrispondenza di entità

CLAIRE per la governance e la conformità dei dati

Intelligenza artificiale e machine learning sono essenziali per automatizzare le attività di data governance più impegnative di oggi: trovare i dati, misurarne la qualità e consentire la collaborazione per governarli. CLAIRE genera automaticamente regole di policy (ad esempio data quality), collega la semantica del business ai metadati tecnici e indirizza gli utenti verso i dati più pertinenti e affidabili secondo le loro esigenze.

Arricchimento automatico della qualità dei dati

CLAIRE utilizza un approccio NLP basato su Stanford NER per analizzare ed estrarre entità da testo non strutturato. In genere, per estrarre entità da stringhe (ad esempio un codice di prodotto), gli utenti finiscono per scrivere regole di analisi utilizzando tabelle di riferimento ed espressioni regolari. Ma quantità, complessità e pattern dei dati sono in costante aumento, e scrivere tutte le regole possibili per abbinare ogni input non è una cosa pratica, né scalabile. Invece CLAIRE utilizza modelli pre-addestrati per identificare ed estrarre entità basate su Stanford NER.

CLAIRE utilizza machine learning per classificare il testo in arrivo, ad esempio Lingua, Tipo di prodotto o Problema con il supporto tecnico. La metodologia di machine learning utilizzata è denominata apprendimento supervisionato con Naïve-Bayes e Max Entropy (regressione logistica multinomiale). L'apprendimento supervisionato viene utilizzato per addestrare modelli e assegnare etichette. Successivamente il modello addestrato può essere distribuito durante l'elaborazione dei dati per etichettare, instradare ed elaborare diverse classi di input, ad esempio trattare i problemi di engine separatamente da quelli di configurazione con significati simili, e distinguere tra usi di parole con significati multipli. CLAIRE automatizza la codifica e la classificazione delle immagini sfruttando i modelli NLP e ML per la classificazione dei prodotti e l'estrazione di meta-tag dalle immagini.

Una grande azienda sanitaria globale impiegava un dipendente a tempo pieno per mappare 21.000 asset tecnici con 6.000 termini di business, un processo che ha richiesto due mesi. Con Axon Data Governance e Enterprise Data Catalog, CLAIRE ha automatizzato la mappatura di 18.000 asset tecnici con una precisione del 99% in 8 minuti.

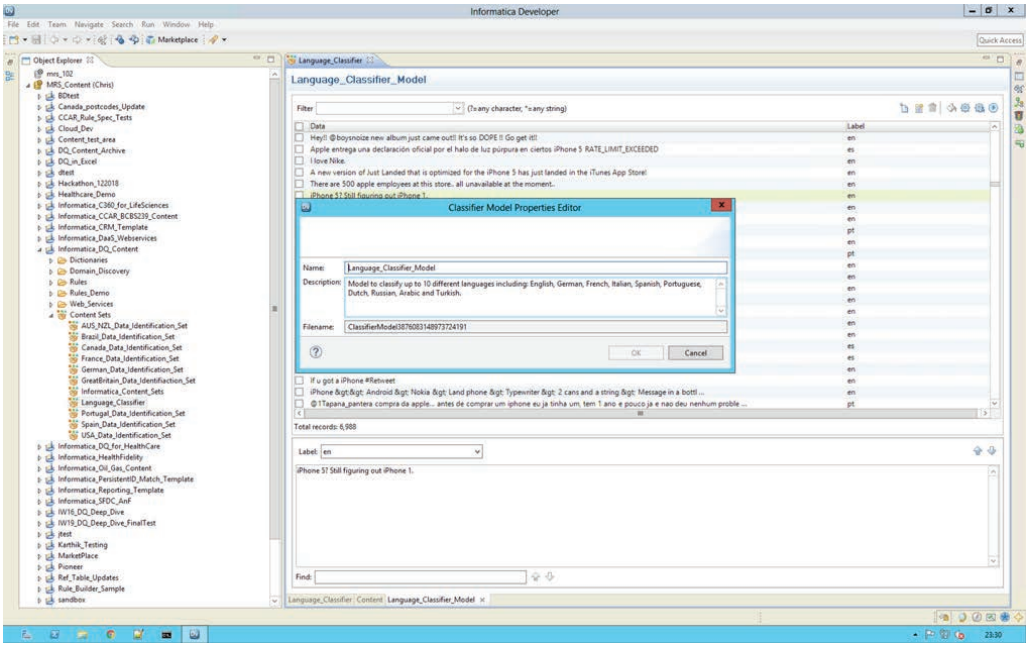

Figura 13. La NLP di machine learning classifica il testo ed estrae le entità.

Associa automaticamente i termini di business ai set di dati fisici

La data governance richiede la documentazione di artefatti di business, definizioni, stakeholder, processi, politiche e altro ancora. Per consentire una vista veramente allineata, è fondamentale che gli utenti siano in grado di associare definizioni e viste aziendali alle implementazioni tecniche sottostanti nel loro patrimonio di dati. In genere, questa attività è lenta, laboriosa e soggetta a errori, poiché si affida a persone chiave per comunicare e allineare manualmente le manifestazioni tecniche una per una, ed è un'attività che può richiedere giorni, settimane o persino mesi.

Informatica Axon Data Governance, attraverso una stretta integrazione con Informatica Enterprise Data Catalog, può abbreviare questo processo. CLAIRE fornisce agli utenti raccomandazioni sugli elementi di dati pertinenti e appropriati da collegare al completamento delle scansioni dei metadati. Ciò riduce il compito di ricerca, convalida e collegamento di elementi di dati, consentendo ai data steward e all'ufficio di data governance di concentrarsi sui propri compiti critici. Man mano che le implementazioni progrediscono, il processo può essere completamente automatizzato.

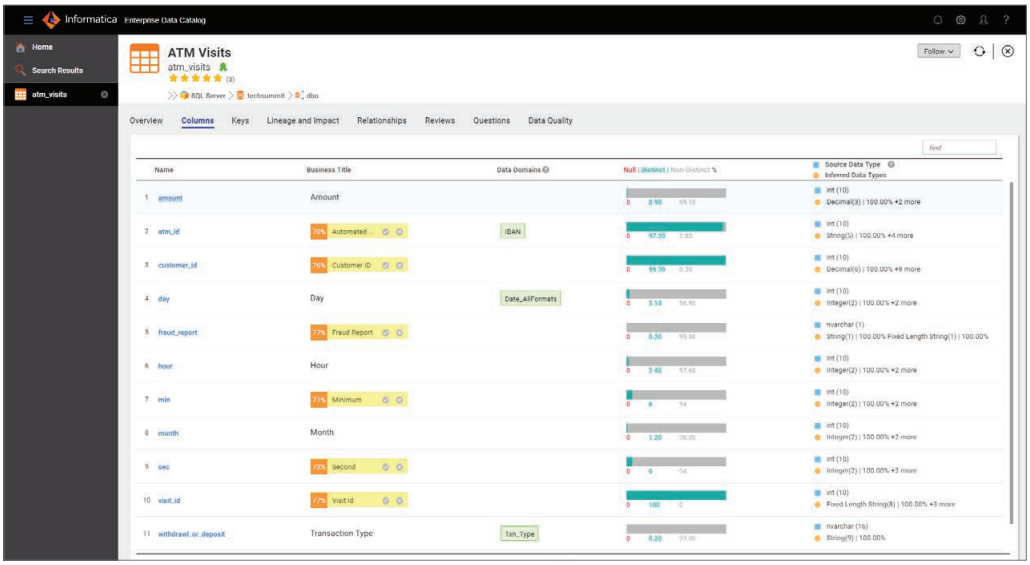

Figura 14. Associazione automatica di termini di business con set di dati fisici.

Valuta automaticamente la qualità dei dati

Un indicatore chiave di prestazione (KPI) nella data governance è la qualità dei dati in un sistema che supporta un processo, sostiene le policy e così via. L'ufficio di data governance deve garantire che i dati siano completi, accurati, coerenti, validi e così via. In breve, deve essere affidabile e abbastanza buono da supportare le attività di business. Man mano che le implementazioni di data governance crescono, la valutazione della qualità per un numero crescente di sistemi e campi nel panorama dei dati, dai database ai data lake, diventa sempre più dispendiosa in termini di tempo.

Attraverso CLAIRE, Axon Data Governance, in coordinamento con Informatica Data Quality e Informatica Enterprise Data Catalog, può automatizzare l'applicazione delle misurazioni della qualità dei dati in tutta l'azienda, facendo risparmiare migliaia di ore di lavoro. Il team di data governance associa ai termini di business e agli elementi di dati critici le regole di qualità dei dati per varie dimensioni di data quality, quindi il sistema sottostante genera i controlli di qualità dei dati richiesti sui vari sistemi e riporta le metriche all'ufficio di governance.

Questa automazione è abilitata combinando tre informazioni chiave:

1. Conoscenza degli elementi aziendali critici e delle regole di qualità dei dati richieste da Axon
2. Regole di qualità dei dati portatili ed eseguibili e un engine di esecuzione flessibile di Informatica Data Quality
3. Dettagli dei metadati da risorse di dati fisici da Enterprise Data Catalog

CLAIRE combina queste informazioni per generare lavori di esecuzione delle regole di data quality in Informatica Data Quality rispetto agli asset di dati fisici da Enterprise Data Catalog. CLAIRE mantiene anche il contesto dell'utente business di Axon per garantire che i risultati vengano visualizzati nelle dashboard corrette e nelle viste aggregate per il consumo da parte dell'ufficio di governance.

L'automazione consente ai programmi di governance di scalare più velocemente che mai, rimuovendo migliaia di ore di lavoro manuale associate alla creazione di valutazioni della qualità dei dati e ricollegandoli al contesto di governance uno per uno. CLAIRE garantisce inoltre che qualsiasi nuovo asset fisico identificato venga automaticamente valutato per la sua qualità. Inoltre, vengono scoperti nuovi domini utilizzando Named Entity Extraction o Classifier nelle regole di qualità dei dati.

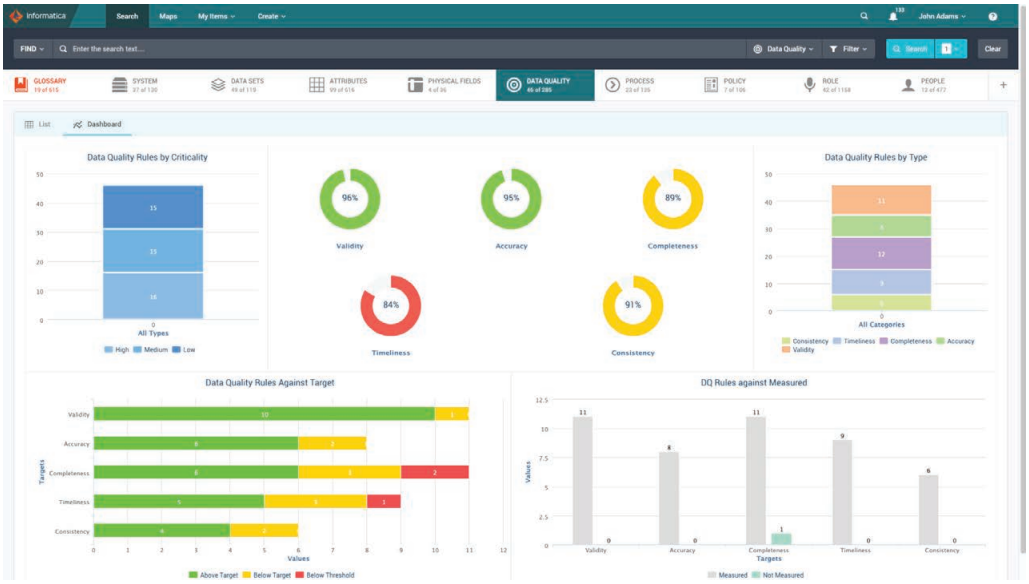

Figura 15. Le valutazioni automatiche della qualità dei dati sull'intero patrimonio di dati consentono di risparmiare migliaia di ore di lavoro manuale.

Regole e identificazione della qualità dei dati assistita da ML/NLP

La qualità dei dati è un obbligo fondamentale per un programma di data governance e nelle implementazioni più grandi possono esserci molte regole sulla qualità dei dati. Per assistere i data steward nell'identificare le regole corrette da utilizzare, CLAIRE può aiutare non solo a identificare le regole, ma anche a generare regole mancanti.

Un utente di Axon Data Governance può specificare il requisito della propria regola in testo normale (ad esempio: "Gli identificatori cliente devono avere otto caratteri e iniziare con C") e invocare l'aiuto di CLAIRE. Attraverso tecniche di ML e NLP, CLAIRE analizzerà il requisito dell'utente e lo tradurrà in una rappresentazione tecnica. Sulla base di questa rappresentazione, nonché dei metadati associati (ad esempio: Nome termine di glossario), CLAIRE cercherà le regole di Informatica Data Quality e identificherà eventuali candidati potenziali. L'utente potrà quindi scegliere da una regola esistente corrispondente o (se non applicabile) richiedere a CLAIRE di generare una nuova regola di data quality.

Se non è stata trovata alcuna regola applicabile, CLAIRE genererà automaticamente una nuova regola di data quality per soddisfare il requisito nel repository di Informatica Data Quality e la collegherà al contesto di Axon Data Governance. Inoltre, CLAIRE assocerà automaticamente le regole di qualità dei dati ai profili Cloud basati su Microsoft Common Data Model (CDM) e fonti Salesforce. Man mano che gli utenti creano nuovi profili rispetto agli oggetti principali da una di queste fonti, CLAIRE suggerirà automaticamente le regole di data quality migliori da applicare alla misurazione.

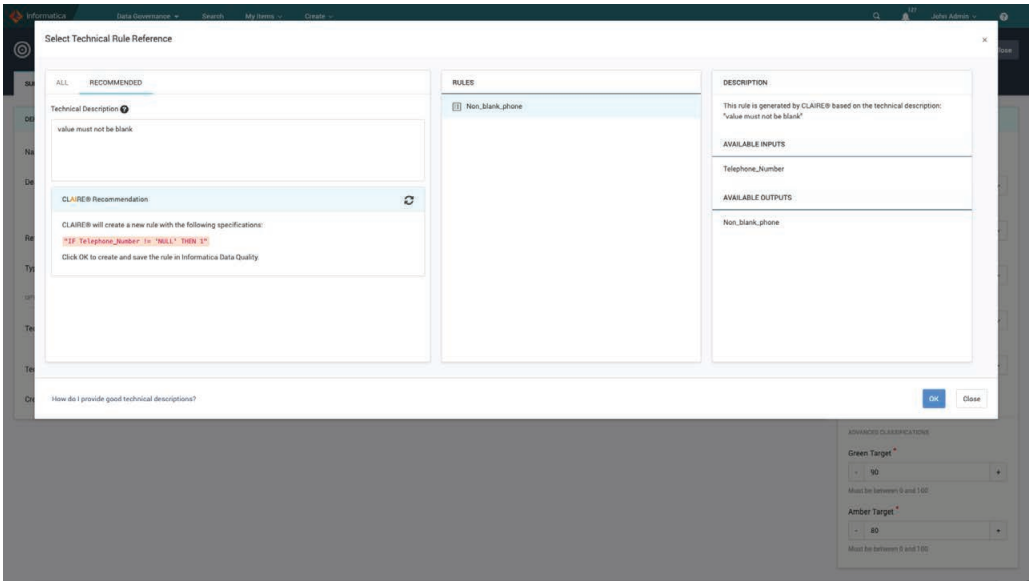

Figura 16. Identificazione automatica delle regole di qualità dei dati utilizzando la NLP.

CLAIRE per la privacy e la protezione dei dati

Con le soluzioni intelligenti per la privacy dei dati fornite da CLAIRE, le organizzazioni possono ottenere una vista e un'analisi a livello aziendale delle informazioni di identificazione personale (PII) all'interno delle risorse di dati. L'automazione basata sull'intelligenza artificiale ti consente di scoprire dati personali e sensibili, comprendere il movimento dei dati, collegare le identità, analizzare i rischi e risolvere i problemi.

Mappatura dell'identità del registro dei soggetti

CLAIRE determina la correlazione dell'identità con i dati sensibili che fornisce la mappatura dei dati per la conformità alla privacy e la segnalazione dell'accesso ai dati. CLAIRE valuta e assegna un punteggio ai dati che in combinazione possono identificare i soggetti dei dati. Oltre alla corrispondenza esatta, vengono utilizzate varie tecniche avanzate, incluso il riconoscimento di entità nominate (NER), per migliorare i risultati comunemente ottenuti quando i dati vengono combinati da diverse fonti.

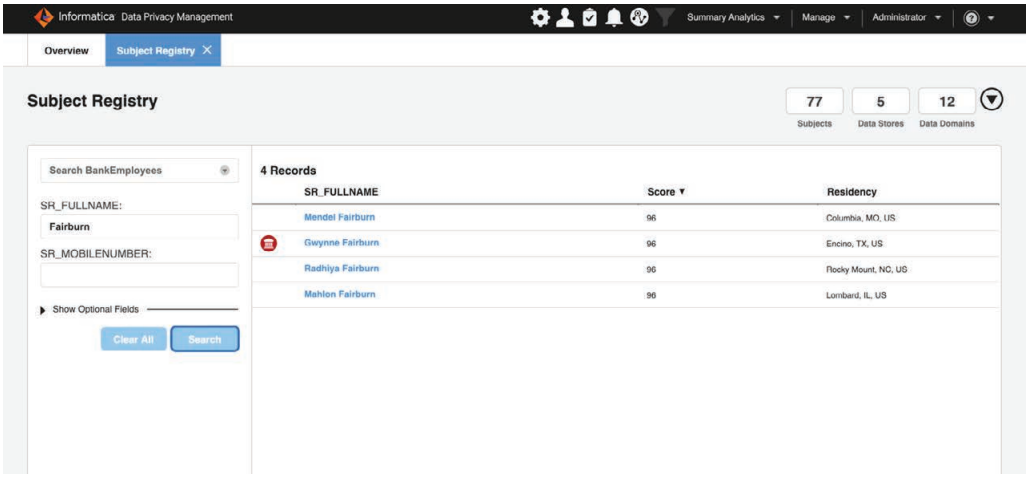

Figura 17. Mappatura dell'identità del registro dei soggetti per la conformità alla privacy e la segnalazione dell'accesso ai dati.

Mappatura e movimento dei dati sensibili

CLAIRE sfrutta ed estende le funzionalità di lineage sopra menzionate per identificare anche il modo in cui i dati sensibili proliferano tra i repository per supportare i requisiti di conformità in materia di sicurezza e privacy. Queste funzionalità possono determinare sia il movimento a monte che a valle, nonché i metadati correlati, come ad esempio il tipo specifico, il processo, lo stato di protezione e la posizione dei dati, per valutare se si sono verificate violazioni. Ad esempio, potrebbe verificarsi una violazione se i dati personali si spostano da una fonte a una destinazione oltre i confini geografici o se i dati inseriti per i processi di fatturazione vengono diffusi in altri dipartimenti o sedi per processi di marketing che potrebbero violare le normative sulla privacy. Gli stakeholder della policy o del processo possono quindi essere informati per apportare la correzione.

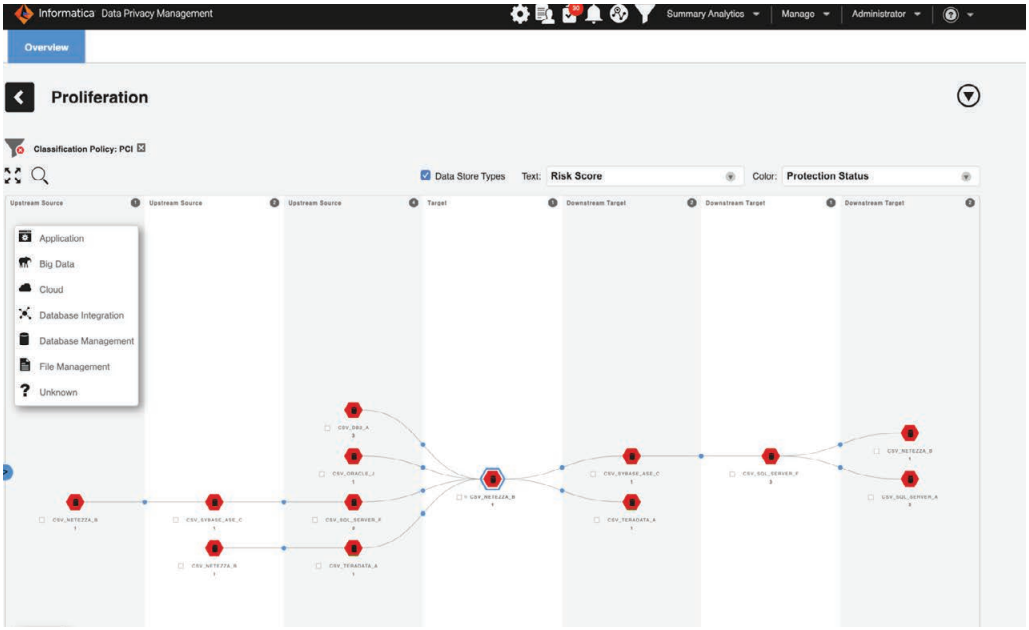

Figura 18. Identificazione e monitoraggio del movimento dei dati sensibili tra i repository.

Piani di simulazione del rischio

Le normative sulla privacy richiedono sempre più che le organizzazioni dispongano di piani di protezione dei dati. CLAIRE può aiutare le organizzazioni a simulare gli impatti di questi piani di protezione per garantire un maggiore ritorno sull'investimento e facilitare i processi di budget. CLAIRE valuta le tecniche di protezione applicate a uno o più domini di dati e quindi calcola la variazione del punteggio di rischio, l'esposizione dei dati sensibili e il costo del rischio residuo per ciascuno degli store di dati selezionati e l'impatto aggregato per l'organizzazione utilizzando un modello di utilità prevista.

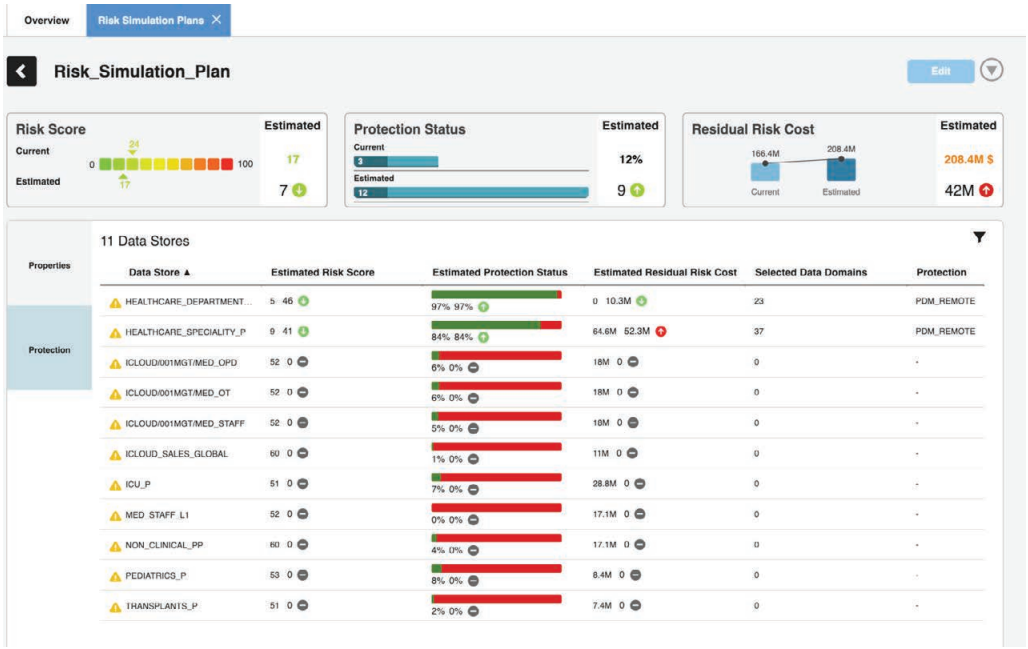

Figura 19. CLAIRE valuta le tecniche di protezione applicate ai domini di dati per determinare il rischio.

Rilevamento intelligente delle anomalie

CLAIRE utilizza approcci statistici e basati su machine learning per rilevare gli outlier e le anomalie dei dati. La funzionalità UBA (User Behavior Analytics) rileva i pattern di comportamento degli utenti che potrebbero essere rischiosi ed esporre un'organizzazione all'utilizzo improprio dei dati. UBA è in grado di rilevare attacchi di tipo impersonation, hijacking delle credenziali e privilege escalation.

UBA applica il machine learning non supervisionato a un modello multi-dimensionale di attività dell'utente, tra cui il numero di store di dati a cui accede l'utente, il numero di richieste eseguite e il numero di record coinvolti nei diversi sistemi. A questo modello viene applicata l'analisi dei componenti principali per la riduzione della dimensionalità. Per il clustering gerarchico non supervisionato viene applicata la tecnica BIRCH al fine di trovare gli utenti che hanno avuto un comportamento diverso durante un determinato periodo di tempo. Per convalidare il comportamento anomalo vengono impiegati metodi di rilevamento degli outlier basati su distanza e densità e viene eseguito il test statistico di Grubbs per gli outlier al fine di confermare che gli oggetti indicati dai primi due metodi siano di fatto outlier nel sistema dei cluster.

Figura 20. Analytics del comportamento degli utenti per rilevare automaticamente le anomalie degli utenti che potrebbero indicare un uso improprio dei dati.

Protezione dei dati nelle API in tempo reale

Proteggi i dati sensibili (ad es. PII) in tempo reale identificando la perdita di dati personali nelle API, bloccando e mascherando i dati. Informatica API Management incorpora librerie di protezione dei dati per bloccare i dati sensibili sulle chiamate API in entrata e in uscita, riducendo al minimo il rischio di esporre dati sensibili.

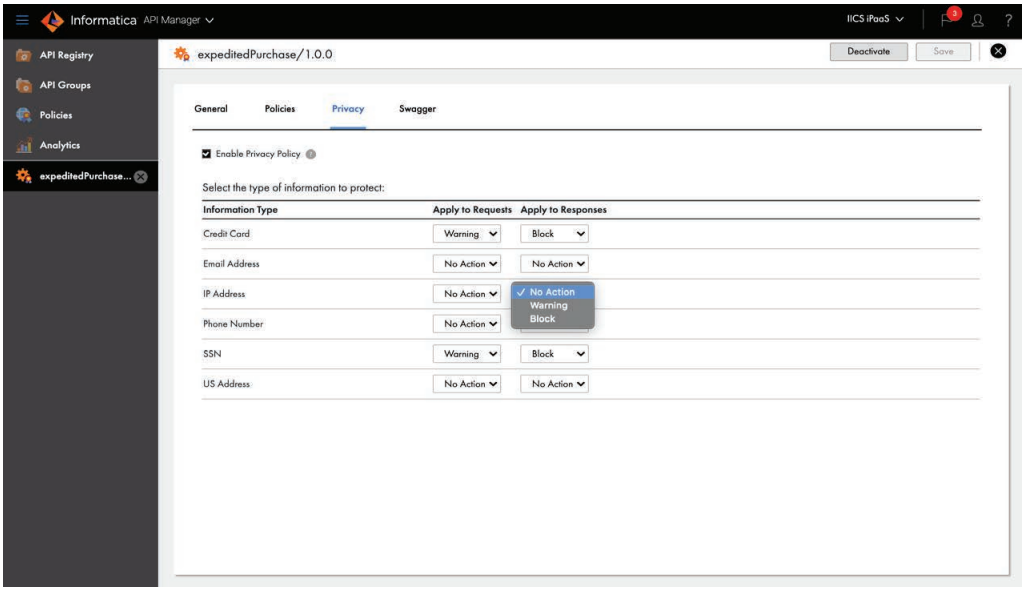

Figura 21. Blocco dell'accesso ai dati sensibili sulle chiamate API in entrata e in uscita.

CLAIRE per DataOps

Con CLAIRE, le organizzazioni possono accelerare le pipeline di elaborazione dei dati, automatizzando molti aspetti della gestione dei dati per l'integrazione continua (CI) e la distribuzione continua (CD) relative a DataOps.

Analytics approfonditi e predittivi per ambienti di gestione dei dati

Gli analytics operativi aiutano a comprendere l'utilizzo corrente di progetti e risorse esistenti e a pianificare la capacità futura. Offre parametri per la creazione di modelli di chargeback supportando più LOB su un'unica piattaforma di gestione dei dati. Sulla base dell'osservazione continua dei trend di utilizzo delle risorse, vengono offerte proiezioni di elaborazione del volume di dati per aiutare con la pianificazione della capacità. CLAIRE consente di fare un passo avanti offrendo il ridimensionamento automatico delle risorse runtime di gestione dei dati.

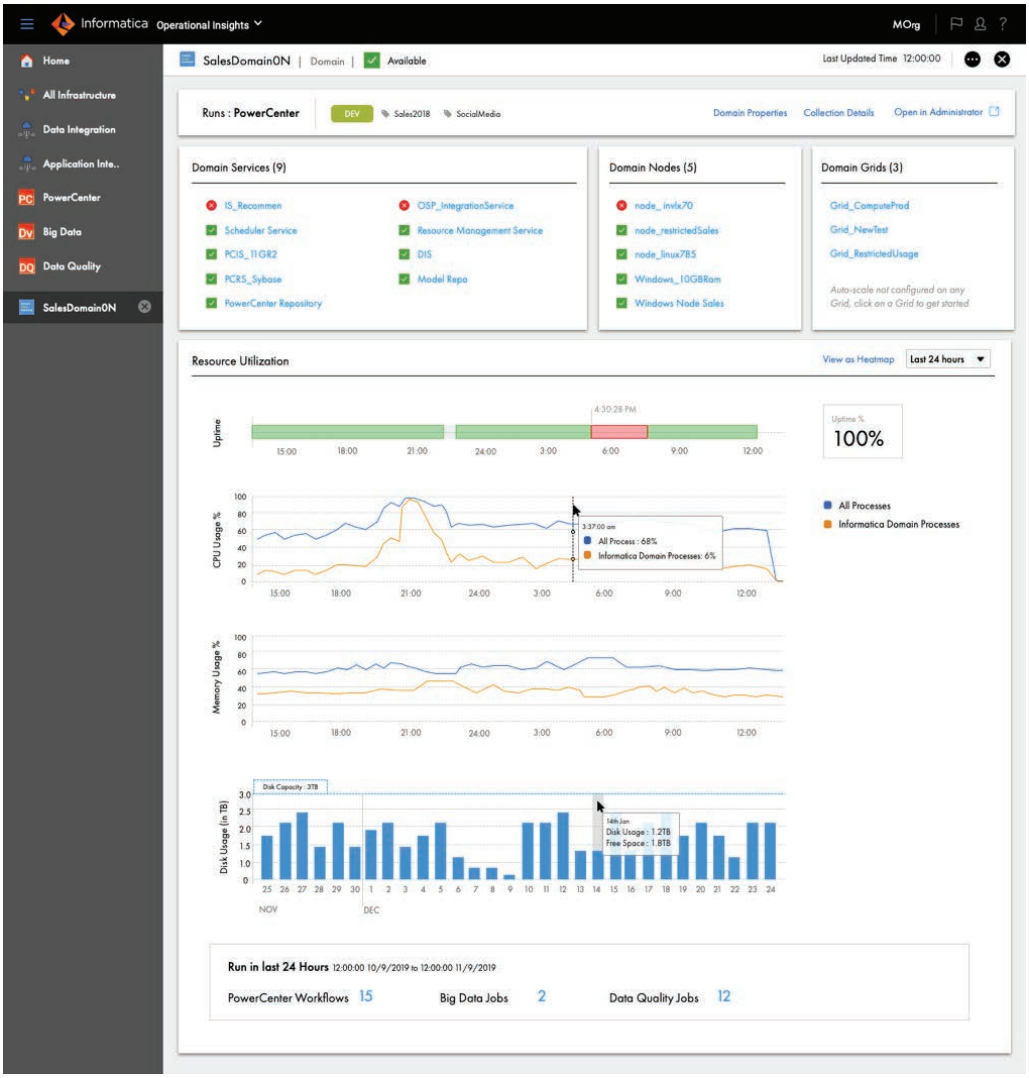

Figura 22. Utilizzo delle risorse di conoscenze operative per i processi di dominio di Informatica.

Rilevamento di anomalie nelle esecuzioni dei lavori

CLAIRE rileva automaticamente le anomalie relative ai tempi di esecuzione dei lavori, ai dati elaborati, ai dati caricati, alle risorse consumate, alla velocità effettiva e altro ancora. Il rilevamento automatico di queste anomalie aiuta l'IT a risolvere in modo proattivo i problemi con i processi di data integration prima che influiscano sui processi di business a valle. L'algoritmo ESD ibrido stagionale viene utilizzato per rilevare anomalie nel comportamento di esecuzione dei lavori. Questo algoritmo prende in considerazione la stagionalità (carico di attività a fine mese, festività natalizie, ecc.) ed elimina i lavori con le aberrazioni previste indotte dai cicli economici.

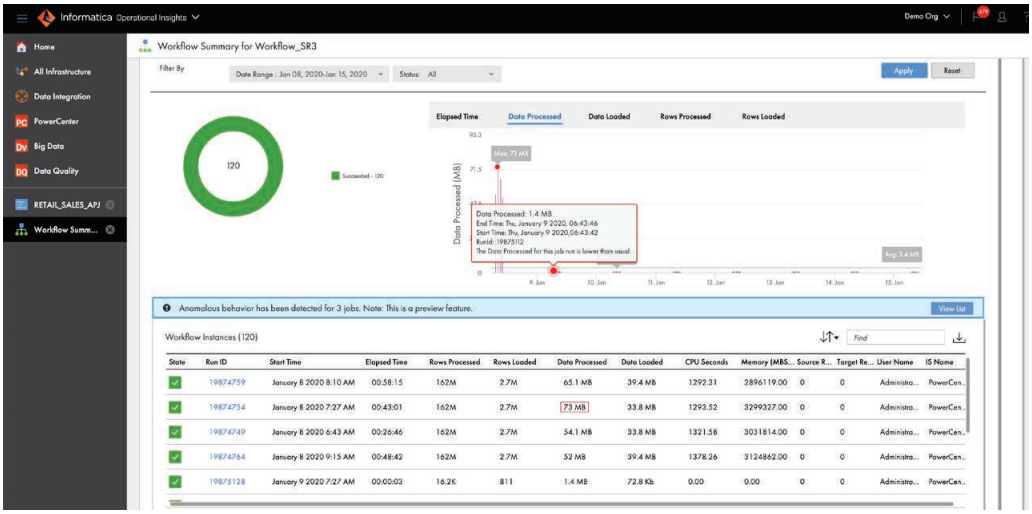

Figura 23. CLAIRE rileva automaticamente le anomalie relative ai job di Informatica e all'elaborazione dei dati.

CLAIRE nel futuro

Man mano che CLAIRE si sviluppa, continuerà ad aumentare la produttività e l'efficienza, consentendo ai leader di dati di sfruttare l'automazione intelligente per informazioni più rapide e migliori e una gestione dei dati più efficace. Le funzionalità future includono:

1. **Integrazione automatica:** per integrare automaticamente i dati nei processi di data integration. Identificazione dei dati, rilevamento dei pattern di integrazione che elaborano dati simili, trasformazione automatica e spostamento dei dati con apprendimento basato su milioni di azioni degli utenti e mappature esistenti.
2. **Assistenza allo sviluppo:** per fornire consigli agli utenti e suggerire le azioni successive nel processo di sviluppo, ad esempio:
   - Completamento automatico delle trasformazioni
   - Consigli relativi ai template
   - Suggerimenti di mascheramento per dati sensibili
   - Consigli di data quality per la bonifica e la standardizzazione
   - Ottimizzazioni automatiche delle performance
3. **Mappatura automatica:** per rilevare le entità di dati master in tutta l'azienda e mapparle automaticamente al modello di dati master applicando trasformazioni dei requisiti e regole di qualità.
4. **Self-healing:** per gestire normalmente i problemi esterni dei sistemi come ad esempio bassi livelli di memoria o potenza di calcolo. Ad esempio, aggiungere potenza di calcolo ("Cloud bursting") per gestire i picchi di dati.
5. **Regolazione automatica:** sulla base delle informazioni storiche, i volumi di dati correnti e le risorse di sistema disponibili prevedono e regolano le pianificazioni o le risorse di calcolo per rispettare i criteri relativi alle performance.
6. **Protezione automatica:** per rilevare automaticamente i dati sensibili e mascherarli prima che lascino un'area sicura.

Conclusione

Le attuali strategie di business focalizzate sui dati sono create su una base composta da dati. Il successo richiede la creazione di competenze di gestione dei dati per liberare la loro potenzialità. Con tutte le sfide che la gestione dei dati presenta in circostanze normali, gli approcci tradizionali non sono in grado di scalare per rispettare i requisiti di oggi, per non parlare di quelli di domani. Un metodo per utilizzare al meglio i dati e promuovere la digital disruption consiste nella standardizzazione di una piattaforma di gestione dei dati end-to-end che utilizzi il potere di dati, metadati e machine learning/AI per migliorare la produttività di tutti gli utenti della piattaforma: tecnici, operativi, business e in particolare self-service del business.

Contattaci per saperne di più su come utilizzare CLAIRE e Intelligent Data Management Cloud per trarre il massimo dalla potenzialità dei dati.

Informatica

Informatica Software Italia srl Piazza della Repubblica 14/16, 20124 Milano Tel: +39 02 37 05 80 00 | Via Luca Gaurico 9/11, 00143 Roma
Tel: +39 06 54 83 21 34 Numero verde negli Stati Uniti: 1.800.653.3871

IN09_0521_03328