databricks

Stato dei dati+AI

La data intelligence e la corsa alla personalizzazione dei LLM

Le organizzazioni puntano sulla democratizzazione di dati e AI

Introduzione

L'AI generativa apre le porte a una nuova era di innovazione, creatività e produttività. Sono passati solo 18 mesi da quando questa tecnologia è entrata nelle conversazioni quotidiane ma le aziende di tutto il mondo stanno già investendo nella GenAI per trasformare la propria organizzazione, con la consapevolezza che i dati sono essenziali per fornire agli utenti un'esperienza di AI generativa di alta qualità. La domanda pressante che si pongono ora i leader di settore è: Qual è il modo più rapido ed efficiente per ottenere questo risultato?

Silos di dati e piattaforme AI rendono difficile per i team accelerare i progetti di AI generativa, ad esempio per interrogare i dati usando il linguaggio naturale o per sviluppare app intelligenti. La diffusione delle piattaforme di data intelligence porterà a una radicale democratizzazione all'interno delle organizzazioni. Questo nuovo tipo di piattaforme utilizza la GenAI per rendere sicuri i dati e sfruttarli più facilmente, riducendo le competenze tecniche necessarie per estrarne valore. Tra i nostri clienti, l'adozione dell'AI è in netta accelerazione.

Stato dei dati + AI fornisce una panoramica delle strategie adottate dalle organizzazioni per dare priorità a iniziative basate su dati e AI. Le informazioni qui condivise provengono da oltre 10.000 clienti di tutto il mondo, inclusi oltre 300 Fortune 500, che utilizzano la Databricks Data Intelligence Platform. Scopri come le organizzazioni più innovative utilizzano con successo il machine learning, adottano la GenAI e rispondono alle esigenze di governance in continua evoluzione.

Questo studio si prefigge di aiutare le aziende a sviluppare strategie di dati efficaci nel mutevole panorama dell'Enterprise AI.

Punti chiave

11 volte di più i modelli di AI messi in produzione quest'anno

Dopo anni di esperimenti, le aziende ora utilizzano un numero significativamente maggiore di modelli nel mondo reale rispetto a un anno fa.

In media, l'efficienza delle organizzazioni nel mettere in produzione i modelli è più che triplicata.

L'elaborazione del linguaggio naturale è l'applicazione di machine learning più usata e in maggiore crescita.

Il 70% delle aziende che ha adottato la GenAI utilizza strumenti e database vettoriali per potenziare i modelli di base

In meno di un anno dalla sua integrazione, LangChain è diventato uno dei prodotti di dati e AI più usati.

Le aziende puntano sulla personalizzazione dei LLM con i propri dati privati usando la Generazione potenziata dal recupero (RAG).

La RAG richiede database vettoriali, cresciuti del 377% su base annua. (Il dato si riferisce LLM sia open source sia chiusi.)

Il 76% delle aziende che fa uso di LLM ha scelto modelli open source, spesso affiancandoli a modelli proprietari

Molte aziende scelgono modelli open source più piccoli per ottimizzare il rapporto tra costo, prestazioni e latenza.

Ad appena 4 settimane dal lancio, Meta Llama 3 rappresenta il 39% dell'utilizzo complessivo di modelli open source.

Sorprendentemente, i primi ad adottare la GenAI sono stati settori altamente regolamentati. Il settore dei servizi finanziari, leader nell'uso di GPU, è quello che si sta muovendo più rapidamente, con una crescita dell'88% in 6 mesi.

Metodologia:
Com'è stato creato questo report Databricks?

Stato dei dati + AI 2024 presenta in forma aggregata e anonimizzata i dati relativi alle modalità con le quali i nostri clienti usano la Databricks Data Intelligence Platform e il suo ampio ecosistema di integrazioni.

Il report si concentra sulle tendenze relative a machine learning, adozione della GenAI, integrazioni e casi d'uso. I clienti inclusi in questo report rappresentano tutti i principali settori industriali e vanno da start-up ad molte delle più grandi aziende del mondo. Se non altrimenti specificato, i dati presentati e analizzati nel report sono stati raccolti dal 1° febbraio 2023 al 31 marzo 2024 e l'utilizzo è misurato per numero di clienti. Laddove possibile, i confronti vengono effettuati su base annua (YoY) per evidenziare i trend di crescita nel tempo.

Machine Learning

L'AI è in produzione

LE ORGANIZZAZIONI ACCELERANO LA MESSA IN PRODUZIONE DEI MODELLI DI ML

Nel corso di quest'anno abbiamo assistito al passaggio dalla sperimentazione sull'AI alla sua applicazione in contesti di produzione. Via via che il machine learning prende piede, le aziende stanno imparando a gestire le due metà del ciclo di vita dei modelli di ML. Le organizzazioni creano i modelli di ML tramite un processo di verifica sperimentale, testando diversi algoritmi e iperparametri; i modelli migliori vengono quindi messi in produzione. In questo processo, i team perseguono due obiettivi contrastanti: velocizzare la fase di sperimentazione e, al tempo stesso, mettere in produzione solo modelli rigidamente testati.

La messa in produzione dei modelli in produzione presenta da sempre numerose sfide: piattaforme dati e AI eterogenee, flussi di lavoro di distribuzione complessi, mancanza di controllo degli accessi per la governance e incapacità di monitoraggio, per citarne solo alcuni. I nostri dati rivelano come le aziende stiano superando queste difficoltà introducendo piattaforme di data intelligence.

Le aziende accelerano i ML in produzione

I dati di MLflow (una piattaforma MLOps open source sviluppata da Databricks) mostrano la frequenza con la quale i nostri clienti caricano modelli (fase di sperimentazione) e li registrano (fase di produzione).

I risultati? Non solo assistiamo a un aumento delle sperimentazioni, ma le aziende stanno anche diventando molto più efficienti nel passare alla produzione.

RAPPORTO TRA ESPERIMENTI CARICATI E MODELLI REGISTRATI

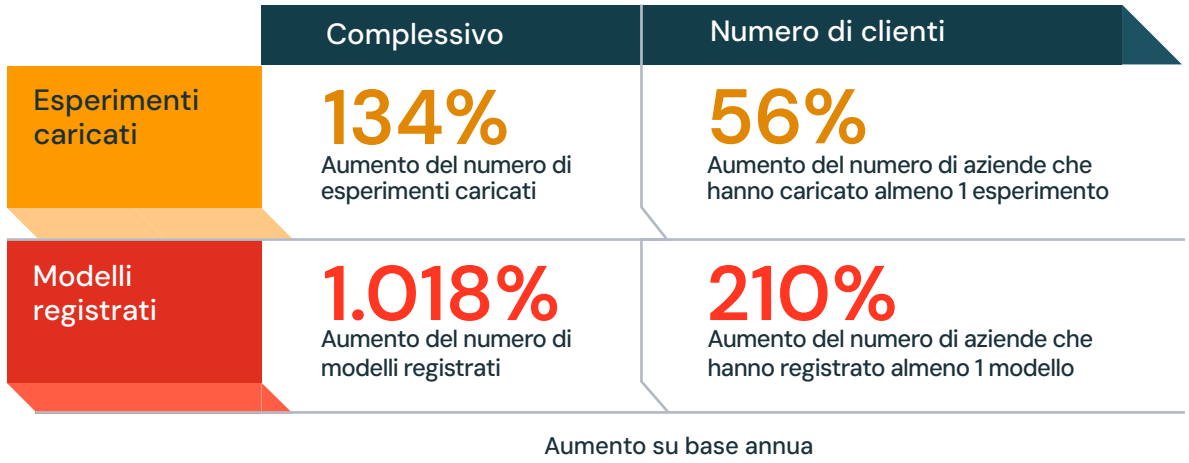
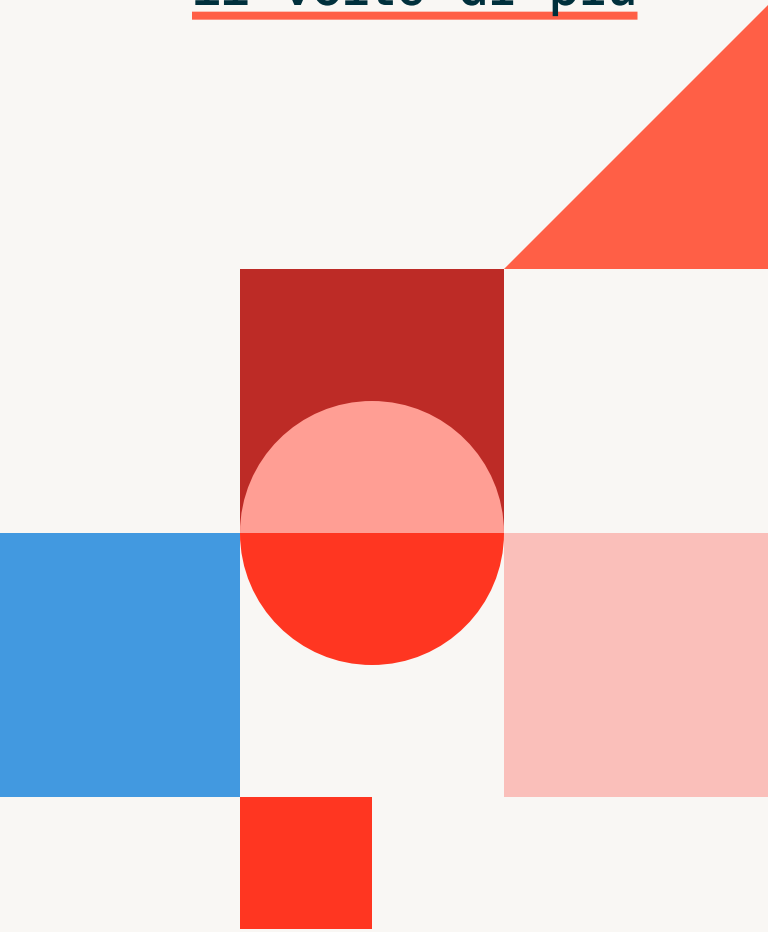


Figura 1:
La crescita su base annua dei modelli registrati ha di gran lunga superato quella degli esperimenti caricati, dunque le aziende stanno passando dalla sperimentazione alla produzione.

Un incremento enorme: i modelli entrati in produzione sono 11 volte di più



Il volume dei modelli è aumentato in modo sostanziale e misurabile.

IL NUMERO DI AZIENDE CHE INVESTONO NEL ML È CRESCIUTO VERTIGINOSAMENTE

I nostri dati mostrano che, rispetto a un anno fa, le aziende che caricano modelli sperimentali sono aumentate del 56%, quelle che li registrano 210%. Questo indica che molte aziende che l'anno scorso si erano concentrate sulla sperimentazione sono passate quest'anno alla produzione.

IL NUMERO DI MODELLI DI ML È AUMENTATO IN TUTTE LE AZIENDE

Dopo anni di intensa sperimentazione, le organizzazioni stanno passando alla produzione. Quest'anno i modelli registrati sono aumentati del 1.018%, a fronte di un aumento del 134% degli esperimenti caricati. Questa tendenza si conferma anche a livello di singole aziende. L'organizzazione media quest'anno ha registrato il 261% di modelli in più e ha caricato il 50% di esperimenti in più.

CONSIDERAZIONI FINALI

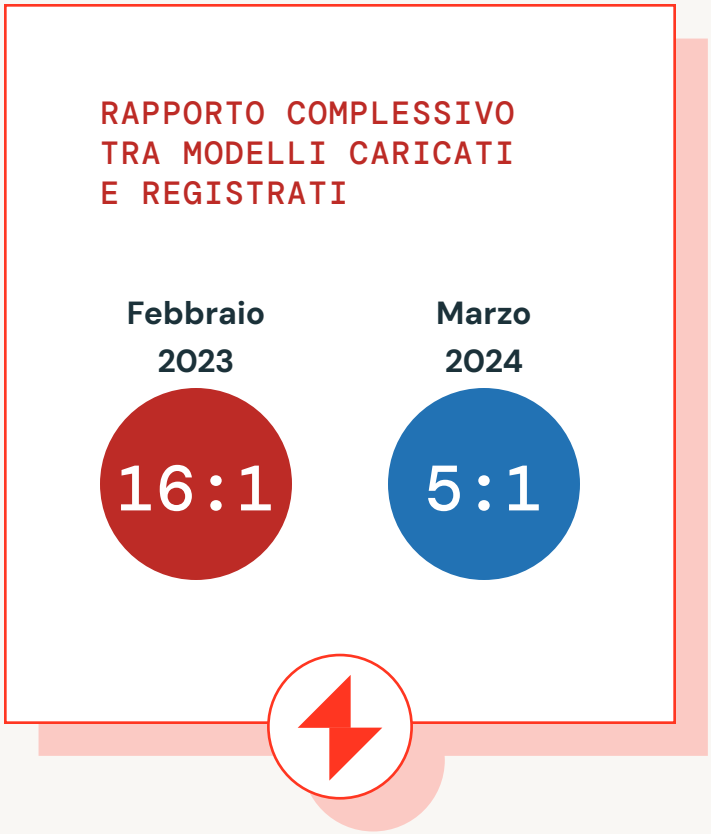
Il ML è un componente essenziale delle strategie di innovazione e differenziazione delle aziende. È prevedibile che, con il progressivo aumento della fiducia nel ML, questa tendenza continui anche nei prossimi anni. Il recente settore della GenAI è ancora in fase di test, ma sta cominciando a trovare interesse nelle aziende.

Le aziende sono diventate 3 volte più efficienti nel mettere in produzione i modelli

L'efficienza del ML ha un valore reale che si può misurare in tempo, denaro e risorse. Per quanto lo sviluppo e la sperimentazione siano fondamentali, è l'implementazione in casi d'uso reali che genera valore per l'azienda.

Per valutare i progressi in questo ambito, abbiamo esaminato il rapporto tra modelli caricati e modelli registrati dai nostri clienti. Nel febbraio 2023, il rapporto tra modelli caricati e modelli registrati era di 16 a 1. In altri termini, per ogni 16 modelli sperimentali, solo uno veniva registrato per la produzione. Entro la fine del periodo preso in esame, il rapporto tra modelli caricati e modelli registrati è sceso a 5 a 1, riducendosi a un terzo.

Cosa ci dicono questi dati? Le aziende metto in produzione i modelli in modo più efficacie, spendendo meno risorse su modelli sperimentali che non generano valore nel mondo reale.



Efficienza a livello di settore

Ogni settore ha specifici set di dati, obiettivi strategici e profili di rischio. Di conseguenza, è prevedibile che le differenze si estendano anche all'approccio al ML, incluso il rapporto tra sperimentazione e produzione.

Per meglio comprendere queste tendenze, abbiamo analizzato sei settori chiave.

RAPPORTO TRA ESPERIMENTI CARICATI E MODELLI REGISTRATI, PER SETTORE

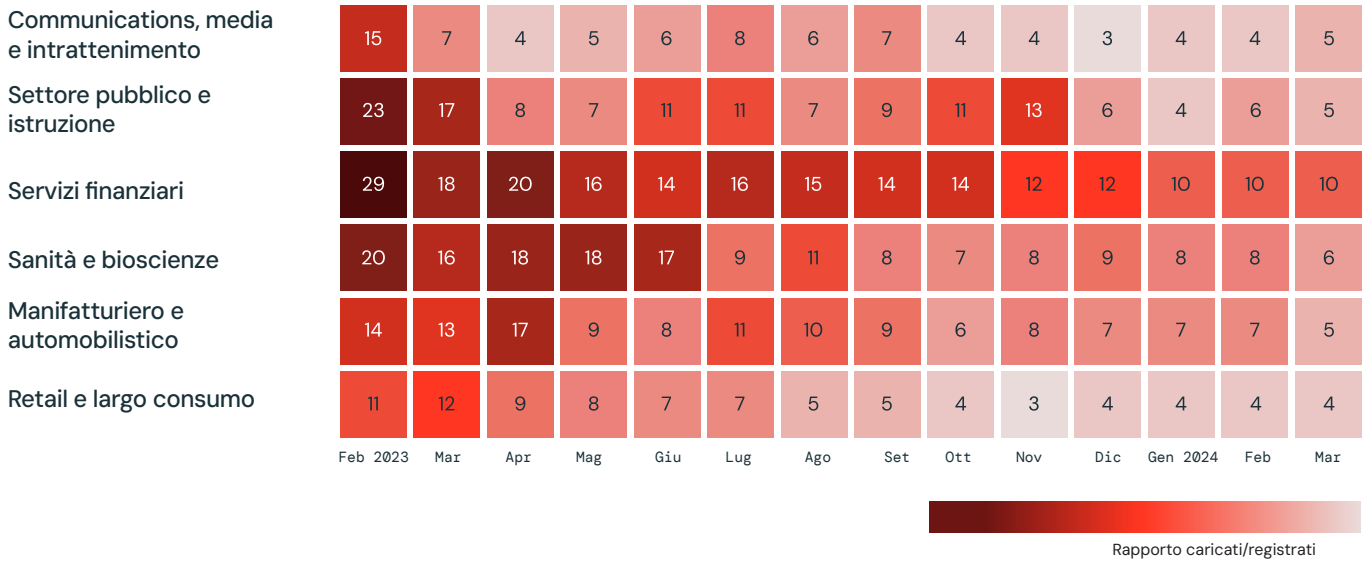


Figura 2:
Il rapporto tra modelli caricati e registrati è andato costantemente calando nel periodo compreso tra il 1° febbraio 2023 e il 31 marzo 2024, dunque le aziende hanno messo in produzione un numero maggiore di modelli sperimentali.

NOTA: A seguito di modifiche nell'API del Model Registry e nel tracciamento, i dati di quest'anno non collimano con il grafico dei modelli caricati e registrati lo scorso anno.

NEL SETTORE PIÙ EFFICIENTE, IL COMMERCIO AL DETTAGLIO, IL 25% DEI MODELLI ARRIVA IN PRODUZIONE

Il settore retail e largo consumo ha raggiunto un rapporto di un modello in produzione per ogni quattro modelli sperimentali, il più efficiente tra tutti i settori in esame. Come indicato nel report MIT Technology Review Insights, il settore retail e largo consumo è da tempo all'avanguardia nell'adozione dell'AI a causa della pressione della concorrenza e delle aspettative dei consumatori.

Aumento di efficienza: il settore dei servizi finanziari è diventato quasi 3 volte più efficiente nel portare i modelli in produzione

IL SETTORE FINANZIARIO REGISTRA L'AUMENTO DI EFFICIENZA PIÙ MARCATO

La sperimentazione risulta preponderante nei servizi finanziari. All'inizio del 2023, nel settore sono stati caricati in media 29 esperimenti per ogni modello registrato. L'efficienza ora è quasi triplicata e nel marzo del 2024 il rapporto è sceso a 10 a 1. Nei settori regolamentati, la posta in gioco per il ML di produzione è più elevata, il che rende essenziali lunghi cicli di test.

Cosa ha permesso quest'anno a varie aziende di mettere in produzione più modelli? Plausibilmente, uno dei fattori determinanti è la disponibilità di piattaforme di data intelligence che forniscono un ambiente aperto e standardizzato per l'intero ciclo di vita del ML. Le aziende possono eseguire tutte le fasi (preparazione dei dati, addestramento del modello, serving e monitoraggio in tempo reale) su un'unica piattaforma, assicurando al tempo stesso governance, privacy e sicurezza. Ciò migliora la qualità dei risultati e accelera la messa in produzione.

Applicazioni di Machine Learning

La crescita vertiginosa dell'NLP

PER IL SECONDO ANNO CONSECUTIVO, L'ELABORAZIONE DEL LINGUAGGIO NATURALE (NLP) SI CONFERMA L'APPLICAZIONE DI DATA SCIENCE E ML PIÙ USATA.

I dati non strutturati sono onnipresenti in tutti i settori e le aree, rendendo essenziale il ricorso a tecniche di elaborazione del linguaggio naturale per interpretarli. La GenAI è un caso d'uso chiave dell'NLP.

La tabella in basso si concentra sulle librerie Python, all'avanguardia nei progressi del ML e dell'AI. Inoltre, Python si conferma regolarmente come uno dei linguaggi di programmazione più popolari. Nei nostri dati, abbiamo aggregato l'uso di librerie Python specializzate per determinare le cinque applicazioni di data science e ML (DS/ML) più utilizzate all'interno delle organizzazioni.

APPLICAZIONI DI DS/ML PIÙ DIFFUSE, PER SETTORE

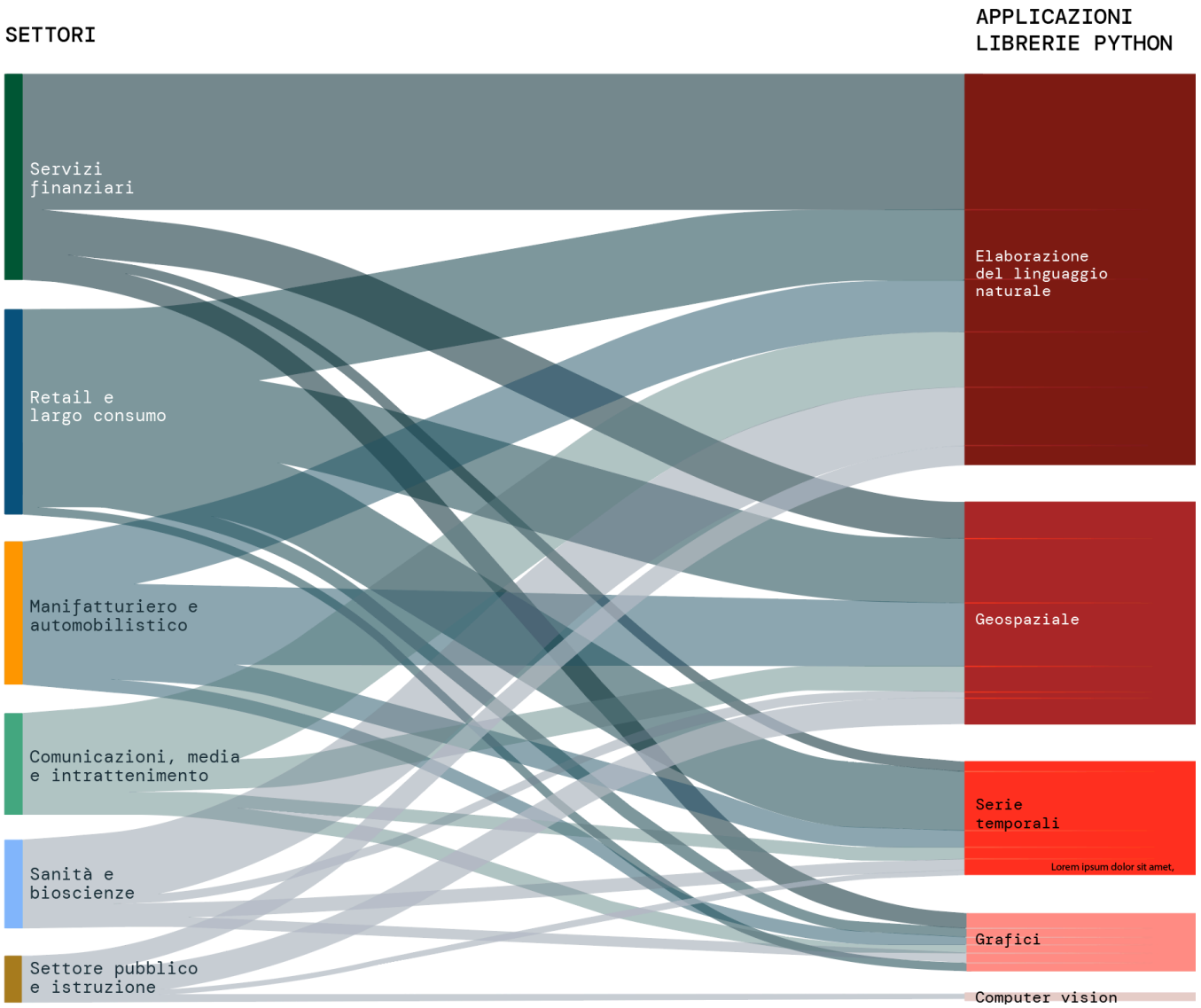


Figura 3:
L'NLP è l'applicazione di libreria Python più usata in tutti i settori oggetto del nostro studio.

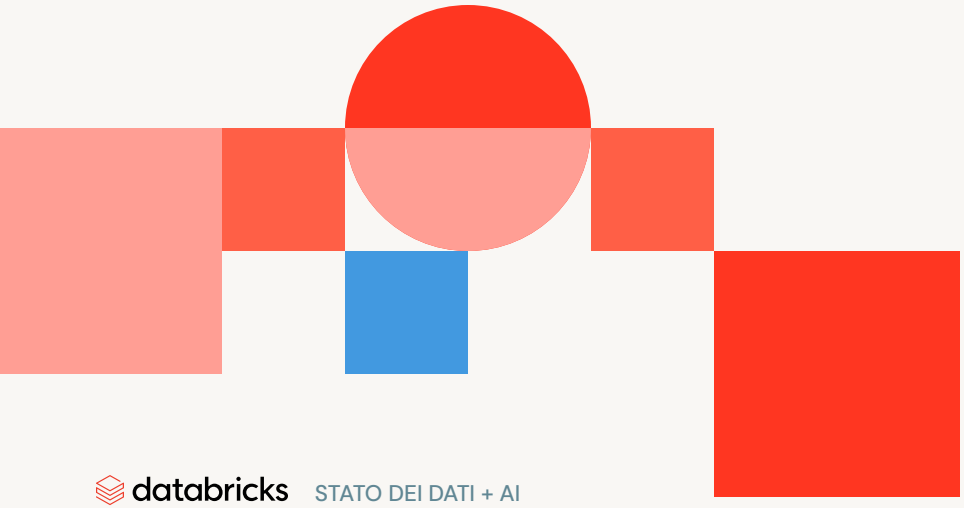
NOTA: Questo grafico mostra il numero univoco di notebook che utilizzano librerie di ML in ciascuna categoria. Non include le librerie usate negli strumenti per la preparazione e la modellazione dei dati.

Per il secondo anno consecutivo, l'NLP si conferma l'applicazione DS/ML più utilizzata; il 50% delle librerie specializzate Python utilizzate sono associate all'NLP.

I team che lavorano sui dati apprezzano inoltre applicazioni geospaziali e di serie temporali. Le librerie geospaziali, spesso usate per analisi basate sulla posizione volte a personalizzare le esperienze degli utenti, sono il secondo caso d'uso per popolarità, e rappresentano il 30% dell'utilizzo complessivo di librerie Python.

L'ADOZIONE DELL'NLP È PIÙ ELEVATA NEL SETTORE SANITÀ E BIOSCIENZE

Tra i settori in esame, quello di sanità e bioscienze ha la più alta percentuale di utilizzo delle librerie Python dedicate all'NPL (69%). Secondo un sondaggio condotto da Arcadia con la Healthcare Information and Management Systems Society, il settore sanitario genera il 30% del volume di dati mondiale e sta crescendo più rapidamente di qualsiasi altro. L'NLP supporta l'analisi nella ricerca clinica, accelera l'immissione di nuovi farmaci sul mercato e aumenta le vendite e l'efficacia commerciale del marketing.

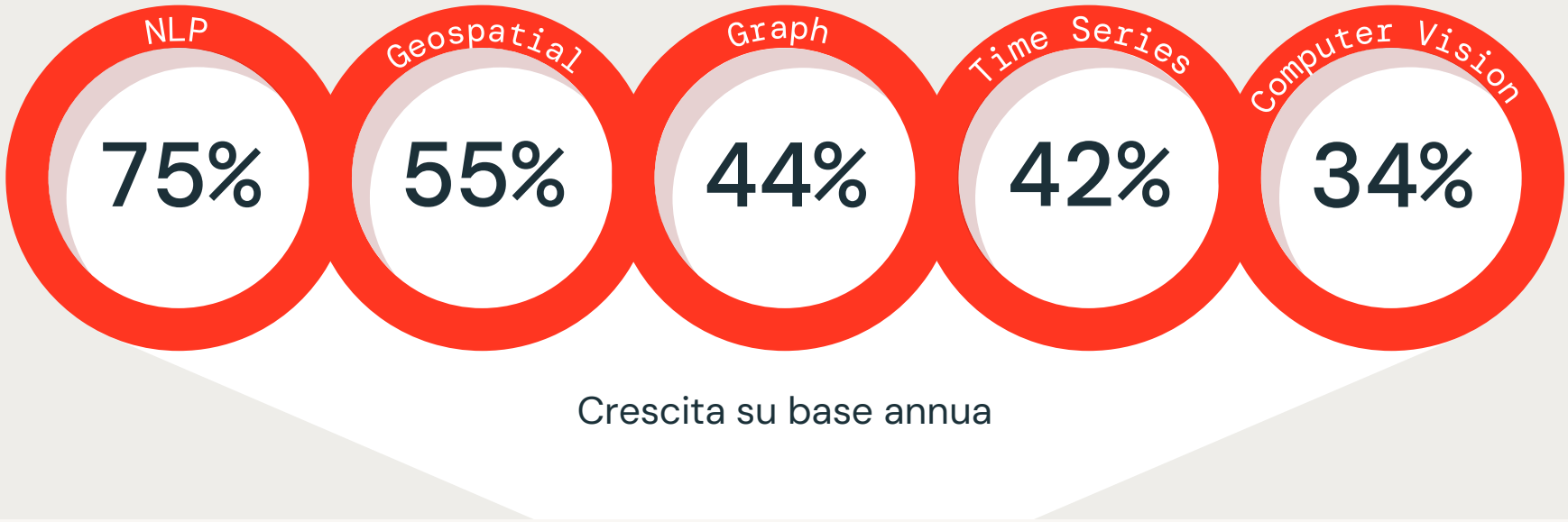


Il 50% delle librerie specializzate Python utilizzate è associato all'NLP

L'NLP, l'applicazione DS/ML più usata, non rallenta la sua corsa

Con il diffondersi di applicazioni alimentate dall'AI, aumenta la richiesta di soluzioni di NPL per i vari settori. Oltre a dominare l'utilizzo delle librerie Python, l'NPL mostra anche la crescita maggiore tra tutte le applicazioni, con il 75% di incremento su base annua.

Applicazioni DS/ML con la crescita più rapida



APPLICAZIONI DI DS/ML CON LA CRESCITA PIÙ RAPIDA, PER SETTORE

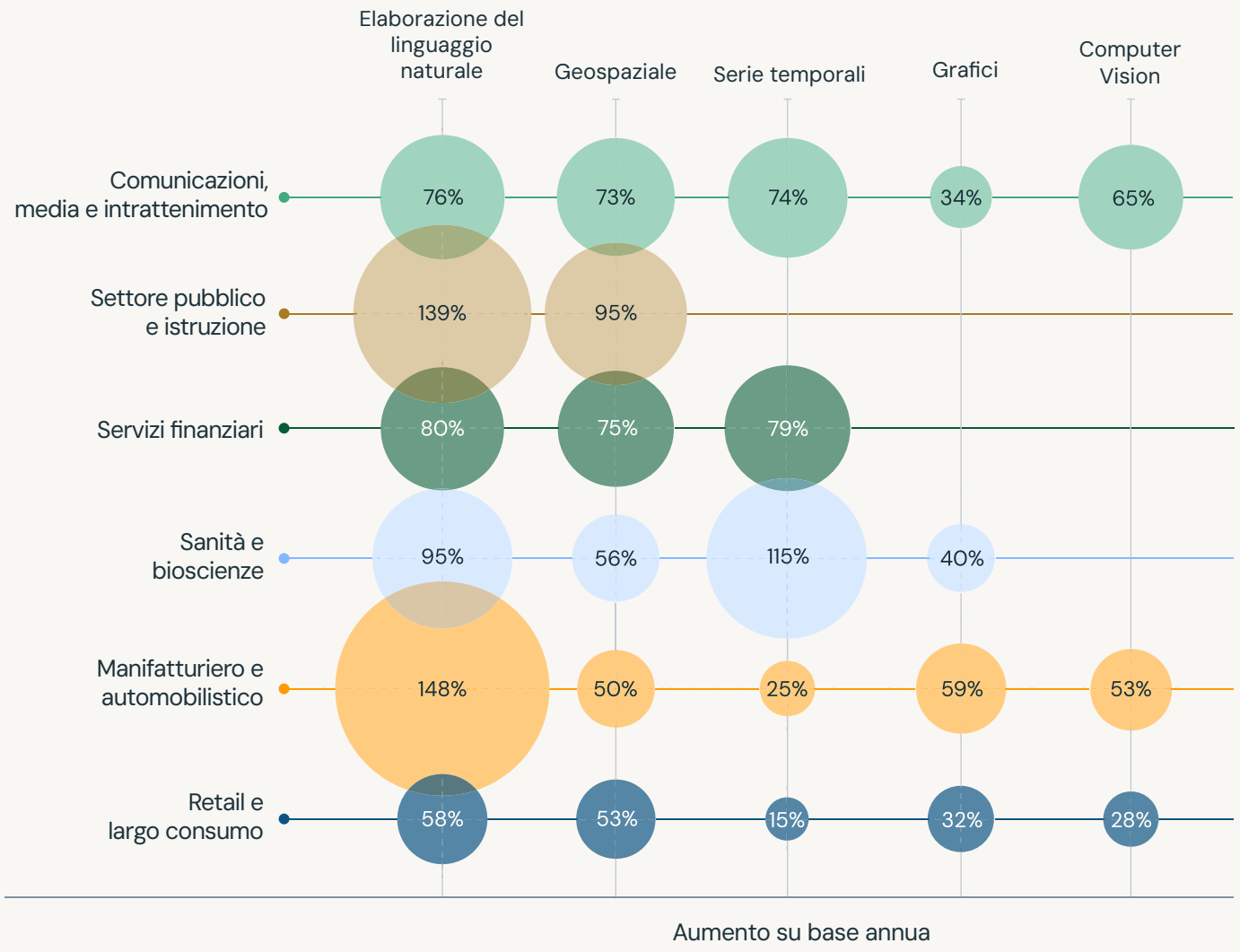


Figura 4:
L'NLP fa registrare la crescita più elevata tra tutte le applicazioni. La crescita maggiore su base annua si registra nel settore manifatturiero e automobilistico, con un incremento del 148% nell'utilizzo dell'NLP.

TUTTI I SETTORI INVESTONO AMPIAMENTE NELL'NLP

Tra i settori presi in esame, quello manifatturiero e automobilistico registra la crescita maggiore nell'utilizzo dell'NLP, con un incremento del 148% su base annua. L'NLP, utilizzata nel settore per un'ampia gamma di applicazioni che vanno dall'analisi dei feedback dei clienti al monitoraggio del controllo qualità e la gestione di chatbot, consente alle aziende di migliorare l'efficienza operativa. Al secondo posto per crescita nell'utilizzo dell'NLP c'è il settore pubblico e dell'istruzione, con un incremento del 139% su base annua.

DAGLI INCENDI BOSCHIVI ALL'INFLUENZA AVIARIA, GLI EVENTI DI ATTUALITÀ COINCIDONO CON LA CRESCITA DEL ML

Le applicazioni geospaziali sono seconde per rapidità di crescita in tutti i settori esaminati. Sempre più spesso, le aziende vanno alla ricerca di schemi, tendenze e correlazioni nei dati basati sulla posizione. L'elevato tasso di crescita delle applicazioni geospaziali nel settore pubblico e dell'istruzione potrebbe essere collegato con la gestione dei disastri e la programmazione dei piani per le emergenze.

L'adozione di librerie di serie temporali nel settore della sanità e bioscienze fa registrare il terzo tasso di crescita più elevato tra tutte le applicazioni e i settori, con un incremento del 115% su base annua. Le serie temporali aiutano le previsioni di rischio per i pazienti, le tendenze di approvvigionamento e la scoperta di farmaci. Uno studio del 2023 condotto dal NIH ha stabilito che "in caso di nuove pandemie, l'analisi delle serie temporali consente di fare previsioni a breve termine in modo più facile e veloce, ricavando stime direttamente dai dati." [1]

1 Applications of Time Series analysis in epidemiology: Literature review and our experience during COVID-19 pandemic, 16 ottobre 2023.

L'evoluzione verso la GenAI

I PRODOTTI DI DATI E AI PIÙ UTILIZZATI RIVELANO LA PROSSIMA FASE DELLA GEN AI

I leader di dati sono sempre alla ricerca degli strumenti migliori per attuare le proprie strategie di AI. I nostri 10 prodotti di dati e AI più utilizzati mostrano le integrazioni più adottate sulla Databricks Data Intelligence Platform. Le categorie includono prodotti DS/ML, governance e sicurezza dei dati, orchestrazione, integrazione e sorgenti di dati.

Tra i prodotti più utilizzati, 9 su 10 sono open source. Le organizzazioni scelgono prodotti più flessibili e meno limitanti rispetto a soluzioni proprietarie. Come vedremo più avanti, si riscontra anche un crescente aumento di popolarità dei LLM aperti.

I 10 PRODOTTI DI DATI E AI PIÙ UTILIZZATI

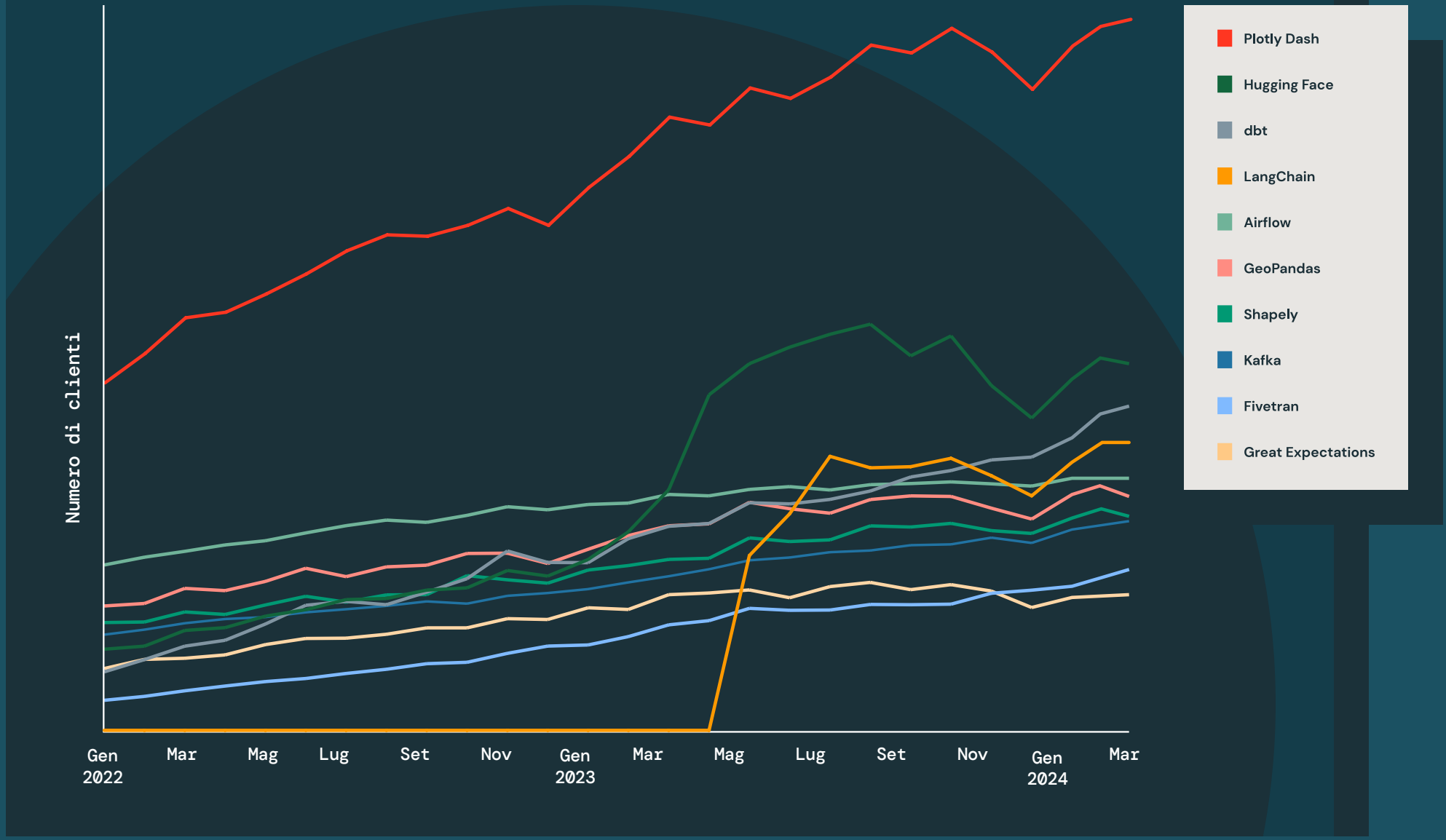


Figura 5: I nostri 10 prodotti di dati e AI più utilizzati coprono le categorie DS/ML, governance e sicurezza dei dati, orchestrazione, integrazione e sorgenti di dati.

PLOTLY DASH MANTIENE IL PRIMATO

Plotly Dash è una piattaforma low-code che consente ai data scientist di creare, scalare e distribuire facilmente applicazioni di dati. Prodotti come Dash aiutano le aziende a fornire applicazioni in modo più facile e veloce per tenere il passo con le esigenze aziendali in continua evoluzione. Da più di 2 anni, Dash resta stabile al primo posto, a testimonianza della crescente pressione sui data scientist per sviluppare applicazioni di dati e AI di tipo produttivo.

HUGGING FACE TRANSFORMERS BALZA IN SECONDA POSIZIONE

Quarto nella classifica dello scorso anno, Hugging Face Transformers sale al secondo posto tra i prodotti più utilizzati dai nostri clienti. Numerose aziende utilizzano i modelli di trasformatori preaddestrati della piattaforma open source, uniti ai dati aziendali, per costruire e ottimizzare i modelli di base. Ciò conferma il trend di crescita riscontrato con le applicazioni RAG.

LANGCHAIN ENTRA IN CLASSIFICA A POCHI MESI DALL'INTEGRAZIONE

LangChain, una toolchain open source per costruire LLM proprietari e lavorare con essi, è entrato in classifica la scorsa primavera ed è salito al quarto posto in meno di un anno dalla sua integrazione. Quando le aziende costruiscono applicazioni LLM moderne e usano librerie specializzate Python di trasformatori per addestrarli, LangChain consente loro di sviluppare interfacce di prompt o integrazioni ad altri sistemi.

LE AZIENDE INVESTONO IN PRODOTTI PER LA COSTRUZIONE DI SET DI DATI DI QUALITÀ

La presenza di tre prodotti per l'integrazione dei dati nella nostra Top 10 indica che la costruzione di set di dati affidabili è una priorità per le aziende. dbt (trasformazione dei dati), Fivetran (automazione delle pipeline) e Great Expectations (gestione della qualità) mostrano tutti una crescita costante. In particolare, dbt ha guadagnato due posizioni rispetto allo scorso anno.

Astro nascente

John Snow LABS

John Snow Labs è un'azienda specializzata in AI e NLP che aiuta le organizzazioni del settore sanità e bioscienze a creare, distribuire e gestire progetti di AI. John Snow Labs consente di aumentare l'accuratezza delle diagnosi, scoprire nuovi farmaci e curare i pazienti tramite l'utilizzo di NLP, modelli ML e GenAI avanzati.

La soluzione di John Snow Labs, pur essendo usata prevalentemente nel settore della sanità e bioscienze, si piazza al quindicesimo posto nella nostra classifica di prodotti di dati e AI. La diffusissima libreria Spark NPL dell'azienda supporta una varietà di task di NPL, inclusi classificazione di testi, riconoscimento di entità e analisti del sentiment, rendendola utile anche in altri settori, come quello dei servizi finanziari.

Database vettoriali

La corsa delle aziende alla personalizzazione dei LLM

I LLM supportano una varietà di casi d'uso aziendali grazie alla capacità di comprensione e generazione di linguaggi. Tuttavia, soprattutto in contesti aziendali, i LLM evidenziano dei limiti. Possono essere fonti inattendibili e tendono a fornire informazioni errate, dette "allucinazioni". In sostanza, i LLM di per sé non conoscono un determinato settore e non sono adatti alle esigenze di una specifica organizzazione.

I nostri dati confermano che un numero crescente di aziende sta passando alla RAG invece di fare affidamento su LLM puri. La RAG consente alle organizzazioni di usare i dati proprietari per personalizzare i LLM e per fornire applicazioni di GenAI di alta qualità. Se si forniscono ai LLM ulteriori informazioni rilevanti, questi tendono a dare risposte più accurate e meno fuorvianti.

La RAG apre la strada alla GenAI in azienda

L'anno scorso, il nostro grafico delle librerie Python per LLM mostrava la rapidissima ascesa dei LLM Saas, cresciuti del 1310% in poco più di 5 mesi. I LLM Saas come GPT-4 sono addestrati su enormi set di dati testuali e sono entrati nel mercato di massa meno di 2 anni fa.

Quest'anno, sono i database vettoriali a dominare il nostro grafico. L'intera categoria dei database vettoriali è cresciuta del 377% su base annua e del 186% se si considera esclusivamente il periodo successivo all'anteprima pubblica di Databricks Vector Search. [2]

CHE COS'È LA RAG?

La Generazione potenziata dal recupero (RAG) è uno schema di applicazione della GenAI che trova dati e documenti rilevanti per una domanda o un'attività e li fornisce come contesto al LLM per favorire la generazione di risposte più accurate.

COME LAVORANO INSIEME DATABASE VETTORIALI E RAG?

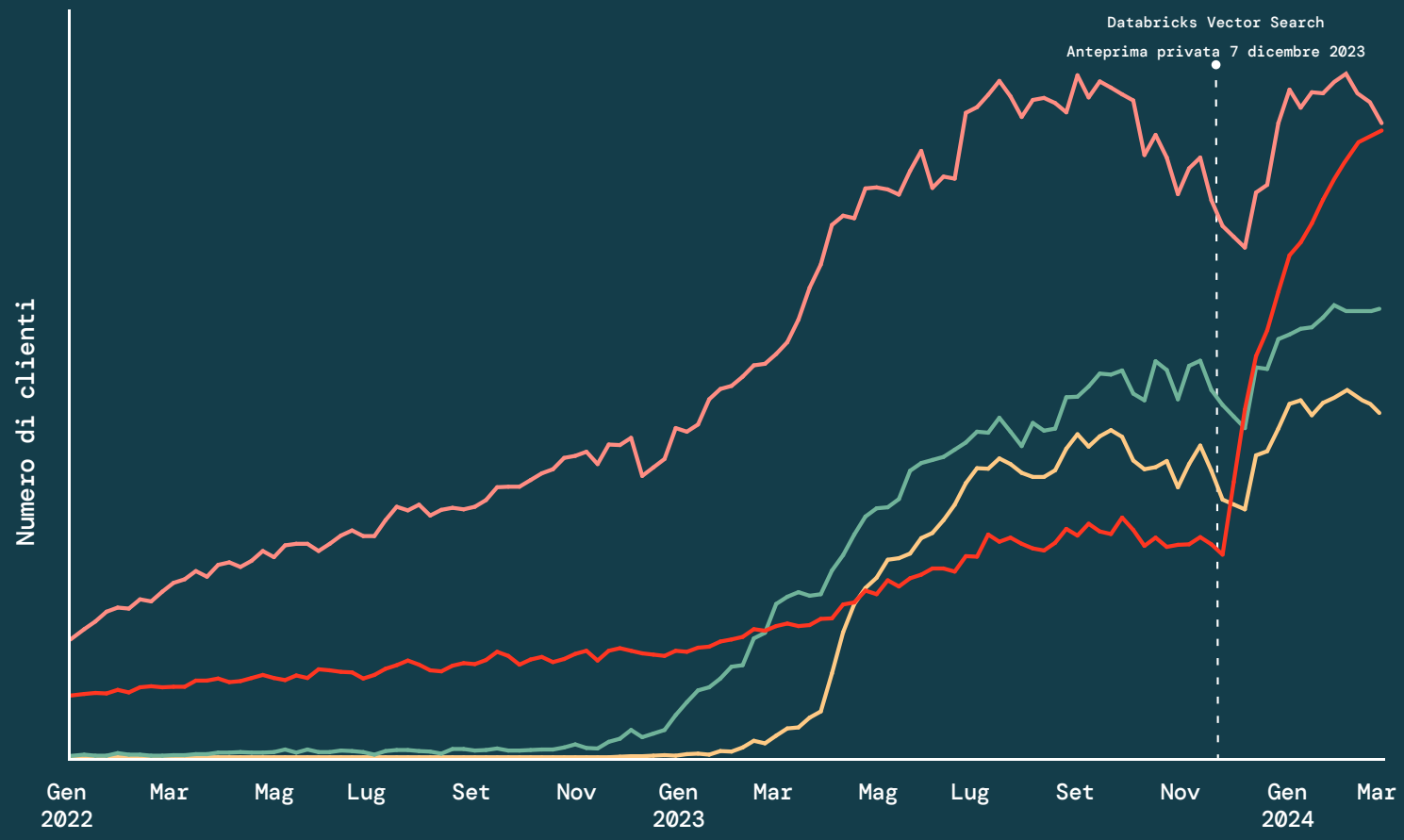
I database vettoriali generano rappresentazioni di dati prevalentemente non strutturati. La loro utilità per il recupero delle informazioni nelle applicazioni RAG sta nel consentire di trovare documenti o record in base alla somiglianza con determinate parole chiave in una query.

Le applicazioni RAG presentano molti vantaggi rispetto ad analoghe soluzioni pronte all'uso. La RAG ha rapidamente preso piede come sistema apprezzato per incorporare dati proprietari in tempo reale nei LLM, risparmiando sui costi e tempi necessari per ottimizzare o preaddestrare un modello.

La crescita esponenziale dei database vettoriali suggerisce che le aziende stiano costruendo più applicazioni RAG per integrare i dati aziendali con i LLM.

2 Databricks Vector Search è stato reso disponibile in anteprima pubblica il 7 dicembre 2023.

USO DELLE LIBRERIE PYTHON PER LLM



Librerie di trasformatori

Database vettoriali

LLM Saas

Strumenti per LLM

LLM: DEFINIZIONI

Addestramento di trasformatori: Librerie per l'addestramento di modelli di trasformatori (es.: Hugging Face Transformers)

LLM Saas: Librerie per accedere a LLM basati su API (es.: OpenAI)

Strumenti per LLM: Toolchain per costruire LLM proprietari e lavorare con essi (es.: LangChain)

Database vettoriali: Indici vettoriali/KNN (es.: Pinecone e Databricks Vector Search)

Figura 6: Dall'anteprima pubblica di Databricks Vector Search, l'intera categoria dei database vettoriali è cresciuta del 186%, molto più di qualsiasi altra libreria Python per LLM.

NOTE: I clienti potrebbero usare più di uno strumento in una data categoria e dunque essere conteggiati più di una volta.
L'utilizzo è misurato in base al consumo di pacchetti che si collegano a servizi di database vettoriali esterni e alle chiamate API sulla nostra piattaforma.
La tendenza tra il 18 dicembre e il 1° gennaio è calcolata su valori medi per tenere conto dei cali stagionali.

LE AZIENDE STANNO ADOTTANDO UN APPROCCIO PIÙ SOFISTICATO ALLA COSTRUZIONE DI LLM.

L'anno scorso i clienti hanno iniziato ad approcciarsi ai LLM con modelli pronti all'uso. I clienti che usano LLM Saas sono ancora in crescita, con un incremento del 178% su base annua. Tuttavia, le aziende cominciano ad assumere maggiore controllo sui propri LLM e a costruire strumenti specifici per le proprie esigenze.

La crescita continua di database vettoriali, strumenti per LLM e librerie di trasformatori dimostra che molti team di dati preferiscono costruire anziché comprare. Le aziende stanno investendo di più negli strumenti per LLM, come LangChain, per costruire LLM proprietari e lavorare con essi. Per addestrare i LLM vengono usate librerie di trasformatori come Hugging Face, che restano di gran lunga le più adottate in termini di numero di clienti. L'utilizzo di queste librerie è cresciuto del 36% su base annua. Considerate nel loro complesso, queste linee di tendenza denotano un approccio più sofisticato all'adozione di LLM open source.

377% di aumento su base annua nel numero di clienti che usano database vettoriali

Le aziende preferiscono i modelli open source di dimensioni minori

USO DI LLM OPEN SOURCE

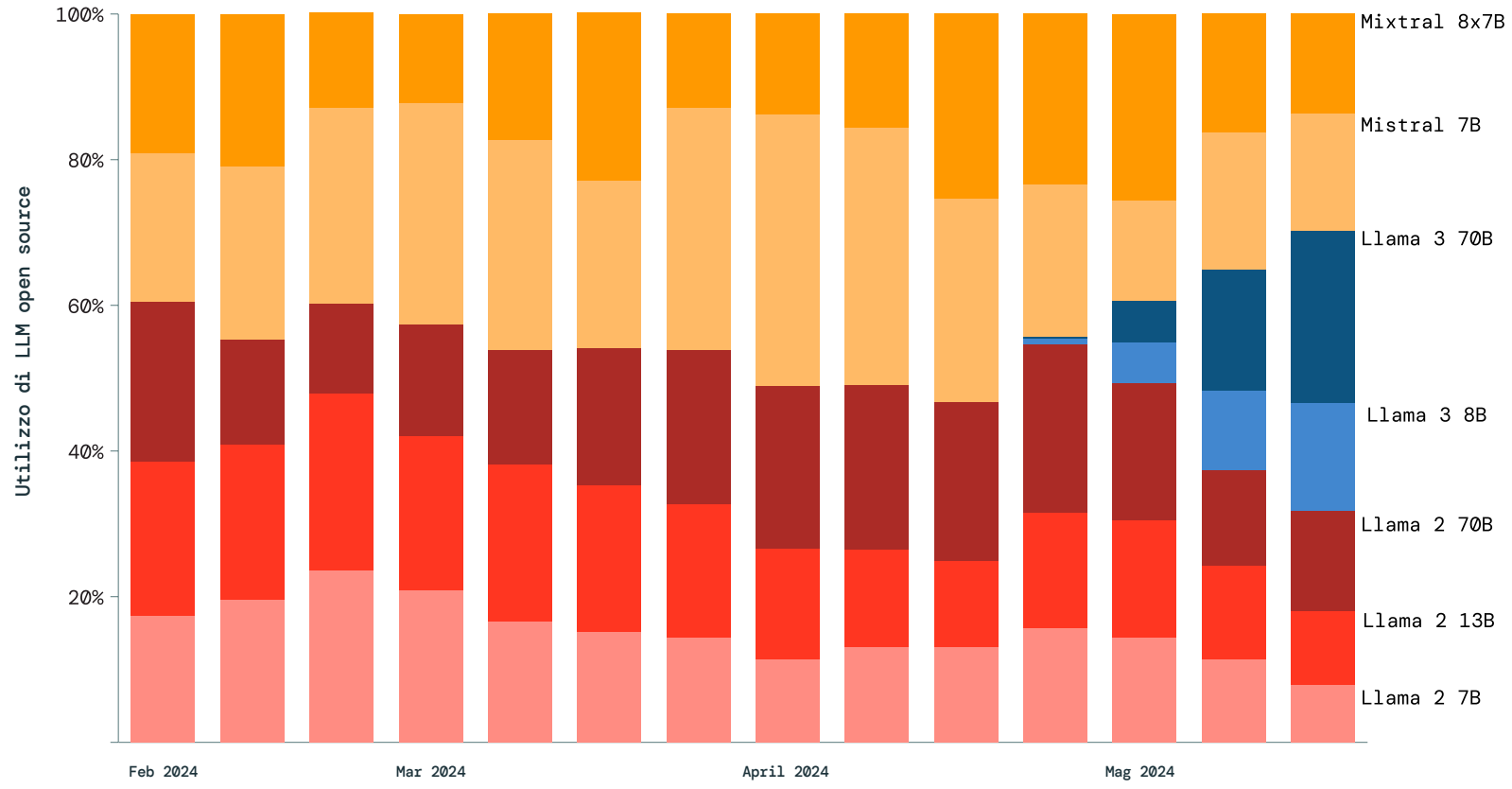


Figura 7: Adozioni rispettive dei modelli open source Mistral e Meta Llama nelle API dei modelli di base di Databricks.

NOTA: Grafico esteso al 19 maggio 2024 per includere il lancio di Meta Llama 3.

Uno dei maggiori vantaggi dei LLM open source è la possibilità di personalizzarli per casi d'uso specifici, soprattutto in contesti aziendali. Spesso ci viene chiesto: Qual è il modello open source più diffuso? In realtà, i clienti spesso provano molti modelli e famiglie di modelli. Abbiamo analizzato l'utilizzo dei modelli open source Meta Llama e Mistral, i due principali. I nostri dati mostrano che il contesto dei LLM aperti è molto fluido e vengono rapidamente adottati nuovi modelli avanzati.

Ogni modello è un compromesso tra costo, latenza e prestazioni. L'utilizzo dei due modelli Meta Llama 2 più piccoli (7B e 13B) è significativamente superiore a quello del modello più grande, Meta Llama 2 70B. Tra gli utenti Meta Llama 2, Llama 3 e Mistral, il 77% ha scelto modelli con massimo 13 miliardi di parametri. Ciò suggerisce che le aziende attribuiscano una notevole importanza a costo e latenza.

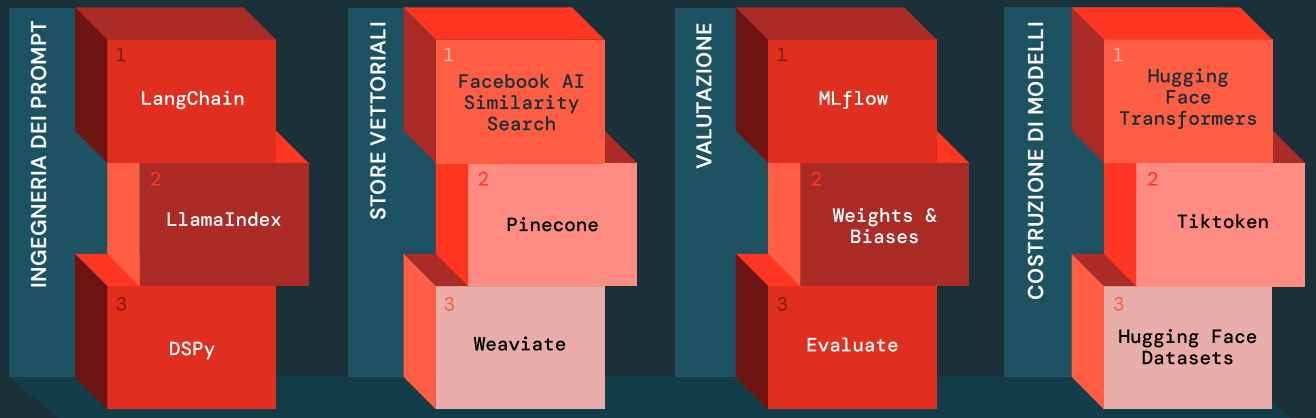
LE AZIENDE PROVANO SUBITO I NUOVI MODELLI

Meta Llama 3 è stato lanciato il 18 aprile 2024. Nel giro di una settimana, aveva già soppiantato altri modelli e fornitori in molte organizzazioni. Dopo appena quattro settimane dal lancio, Llama 3 rappresentava il 39% dell'utilizzo complessivo di tutti i LLM open source.

76%
delle aziende che fa uso di LLM sceglie modelli open source, spesso affiancandoli a modelli proprietari.

70%
delle aziende che ha adottato la GenAI utilizza strumenti, RAG e database vettoriali per personalizzare i modelli.

Migliori pacchetti Python per la GenAI



AI generativa

I settori altamente regolamentati sono i primi utilizzatori

I settori altamente regolamentati sono ritenuti poco inclini ai rischi correlato all'adozione di nuove tecnologie. Le ragioni sono molteplici: rigidi requisiti di conformità, sistemi legacy radicati costosi da sostituire e la necessità di approvazione normativa prima dell'implementazione.

Tutti i settori stanno accogliendo le innovazioni introdotte nell'AI, ma due ambiti altamente regolamentati (servizi finanziari e sanità e bioscienze) mantengono il passo con le loro controparti meno regolamentate e spesso addirittura li sorpassano.

Nel dicembre 2023, Databricks ha rilasciato API per modelli di base, fornendo accesso immediato a popolari LLM open source, come i modelli Meta Llama e MPT. Ci aspettiamo che l'interesse verso l'open source aumenti in maniera significativa parallelamente al rapido miglioramento dei modelli, come mostrato dai recenti lanci di Llama 3.

SFRUTTARE LLM APERTI PER ESIGENZE SPECIFICHE DI SETTORE

I settori manifatturiero e automobilistico e quello di sanità e bioscienze sono al primo posto per quanto riguarda l'adozione di API per modelli di base, e registrano l'utilizzo medio più elevato per cliente. Nell'industria manifatturiera, l'ottimizzazione della supply chain, il controllo della qualità e l'efficienza sono considerati i casi d'uso più promettenti.

Un recente report di MIT Tech Review Insights riferisce che, tra tutti gli intervistati, i CIO del settore sanità e bioscienze ritengono che la GenAI porterà valore alle loro organizzazioni. I LLM open source permettono a settori altamente regolamentati come quello della sanità di integrare la GenAI mantenendo al tempo stesso il massimo controllo sui loro dati.

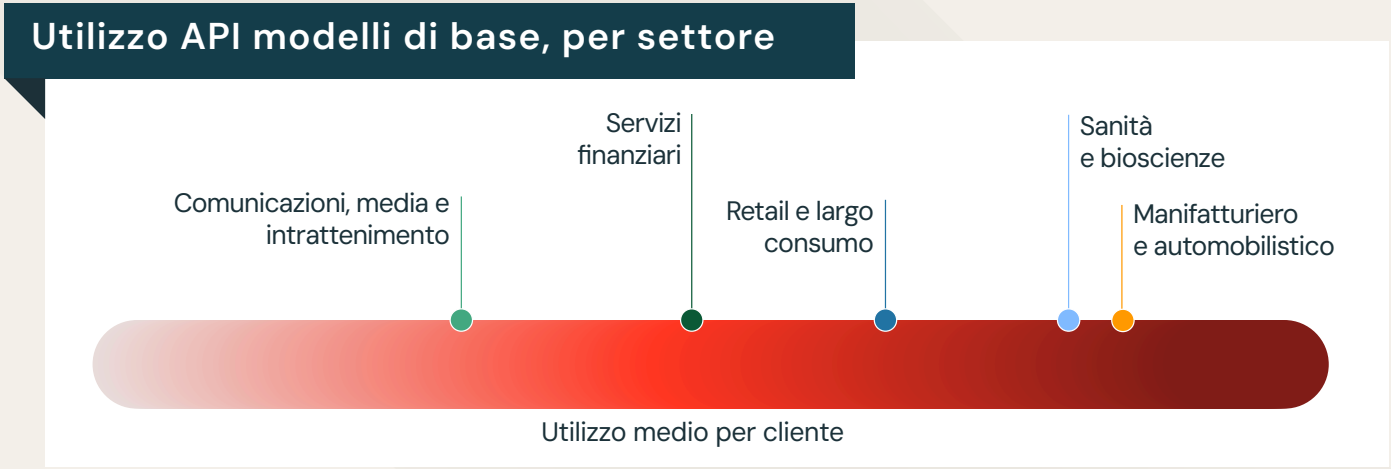


Figura 8: I settori manifatturiero e automobilistico e quello di sanità e bioscienze sono al primo posto nell'adozione di API per modelli di base, con l'utilizzo medio più elevato per cliente.

NOTA: Periodo di riferimento: gennaio - marzo 2024.

CPU vs. GPU: L'adozione di LLM nei servizi finanziari aumenta dell'88% in 6 mesi

Le CPU sono processori general-purpose progettati per gestire rapidamente un'ampia gamma di compiti, ma ne possono gestire un numero limitato in parallelo e vengono utilizzate per il ML classico. Le GPU sono processori specializzati che possono elaborare in parallelo migliaia o milioni di compiti distinti e sono necessarie per addestrare e distribuire LLM.

Abbiamo analizzato l'utilizzo e la crescita di CPU e GPU tra i nostri clienti che usano il Model Serving per comprendere il loro approccio all'AI. Secondo i nostri dati, le GPU sono prevalentemente associate ai LLM.

IL SETTORE FINANZIARIO PRIMO NELL'UTILIZZO DELLE GPU

Uno dei settori più regolamentati, quello dei servizi finanziari, ha adottato con fiducia la GenAI. Il settore ha di gran lunga il più alto utilizzo medio di GPU per azienda, nonché la più alta crescita di GPU, con un incremento dell'88% negli ultimi 6 mesi. I LLM supportano casi d'uso business-critical, inclusi rilevamento di frodi, gestione patrimoniale e applicazioni per investitori e analisti.

MODEL SERVING: ML CLASSICO VS. LLM

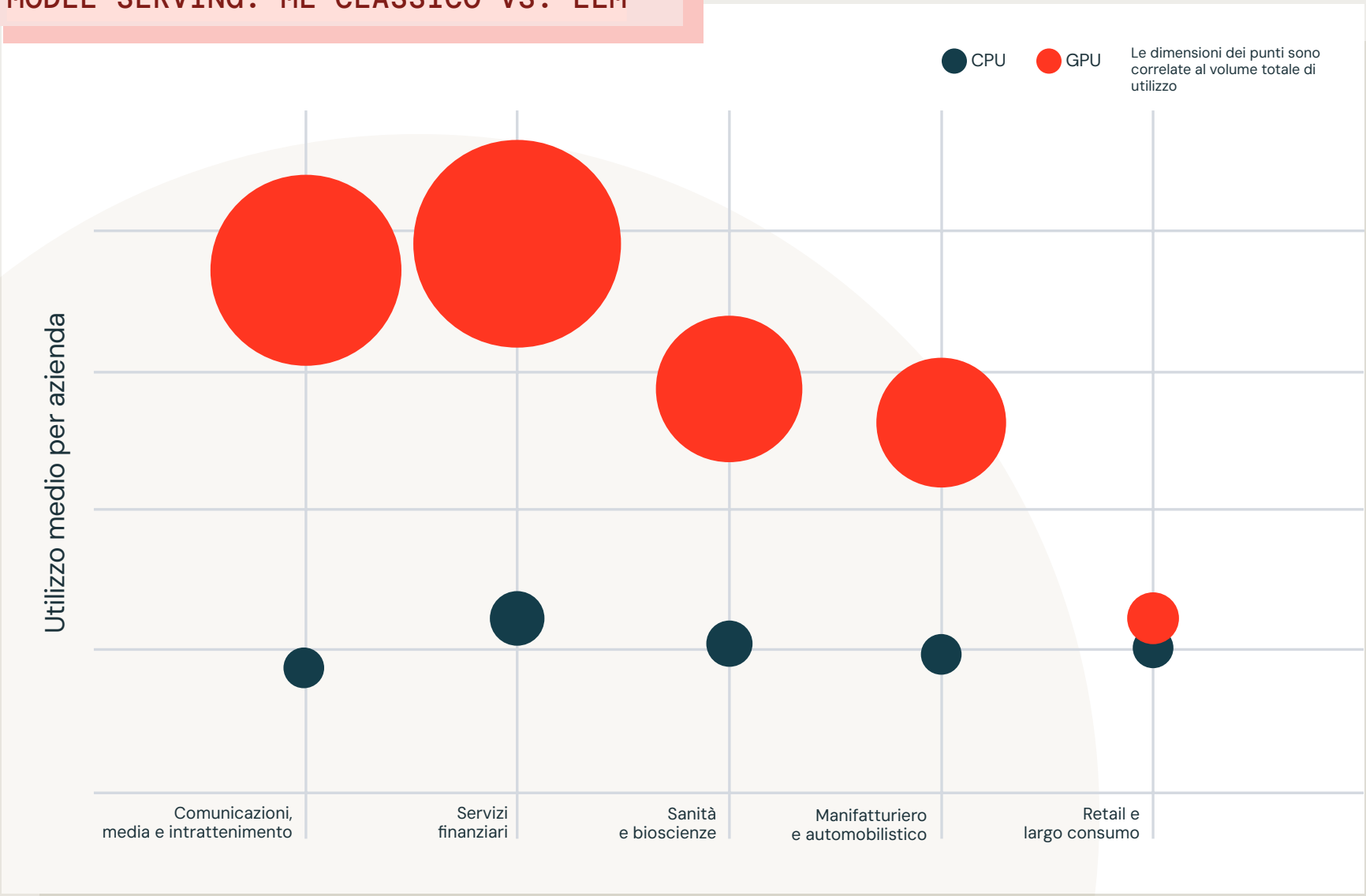


Figura 9: Il settore dei Servizi finanziari ha il più alto utilizzo medio di CPU e GPU.

NOTA: Periodo di riferimento: gennaio – marzo 2024.

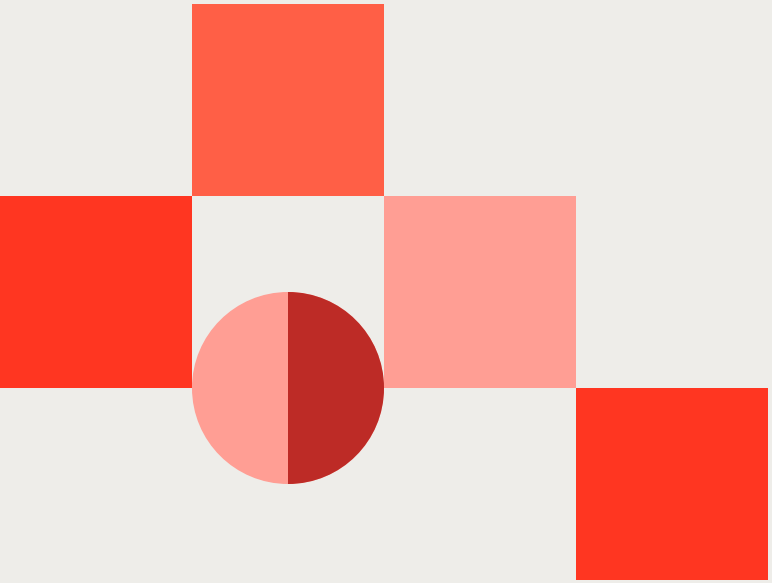
I settori altamente regolamentati primi nell'adozione di una governance unificata

Sicurezza dell'AI e governance sono essenziali per generare fiducia nelle iniziative di AI di un'organizzazione e aiutano i professionisti dei dati a sviluppare e gestire i prodotti nel rispetto di linee guida e standard precisi. Le soluzioni di governance unificata, come Databricks Unity Catalog, coprono tutti i dati e le risorse AI, rendendo più semplice addestrare e distribuire modelli di GenAI sui dati privati di un'organizzazione.

Secondo Gartner, fiducia nell'AI e gestione di rischio e sicurezza sono le tendenze che influenzeranno maggiormente le scelte commerciali e tecnologiche nel 2024. Ora più che mai, i leader vogliono poter sfruttare dati e AI per trasformare le proprie organizzazioni. Lo conferma la diffusa adozione di soluzioni di governance unificata tra i nostri clienti.

IL SETTORE FINANZIARIO ALL'AVANGUARDIA NELLA GOVERNANCE DI DATI E AI

Il rispetto di requisiti di conformità normativa e di sicurezza è un aspetto radicato nella cultura delle aziende operanti nel settore finanziario. Stando ai dati del sondaggio del CIO Vision 2025 Report di MIT Technology Review Insights, nelle istituzioni finanziarie gli investimenti per la gestione dei dati e dell'infrastruttura aumenteranno del "74% da qui al 2025, secondo le stime degli intervistati del settore, rispetto al 52% previsto dal campione nel suo complesso."



ADOZIONE DI UNITY CATALOG, PER SETTORE

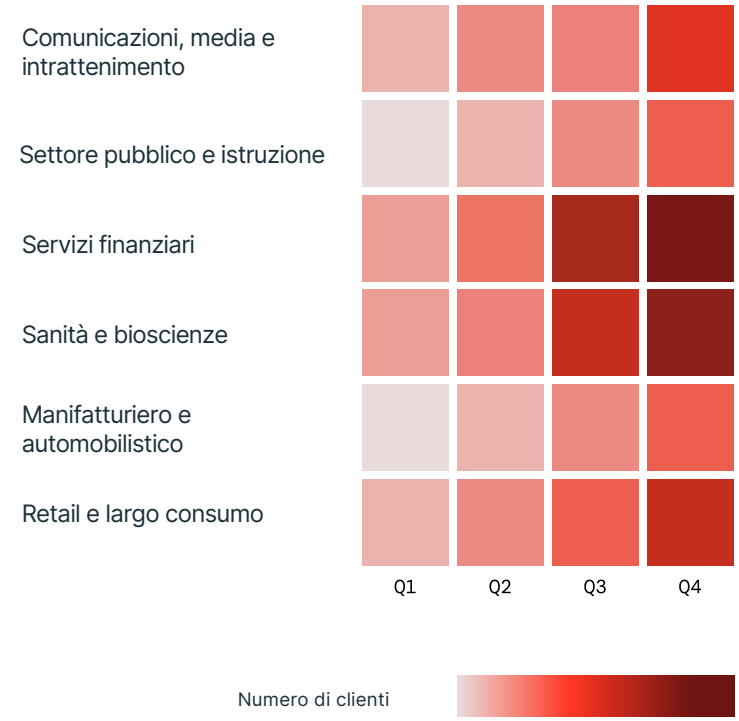


Figura 10: Il settore finanziario primo nell'adozione di Unity Catalog per la governance unificata di dati e AI.

NOTA: Periodo di riferimento: 1° febbraio 2023 – 31 gennaio 2024.

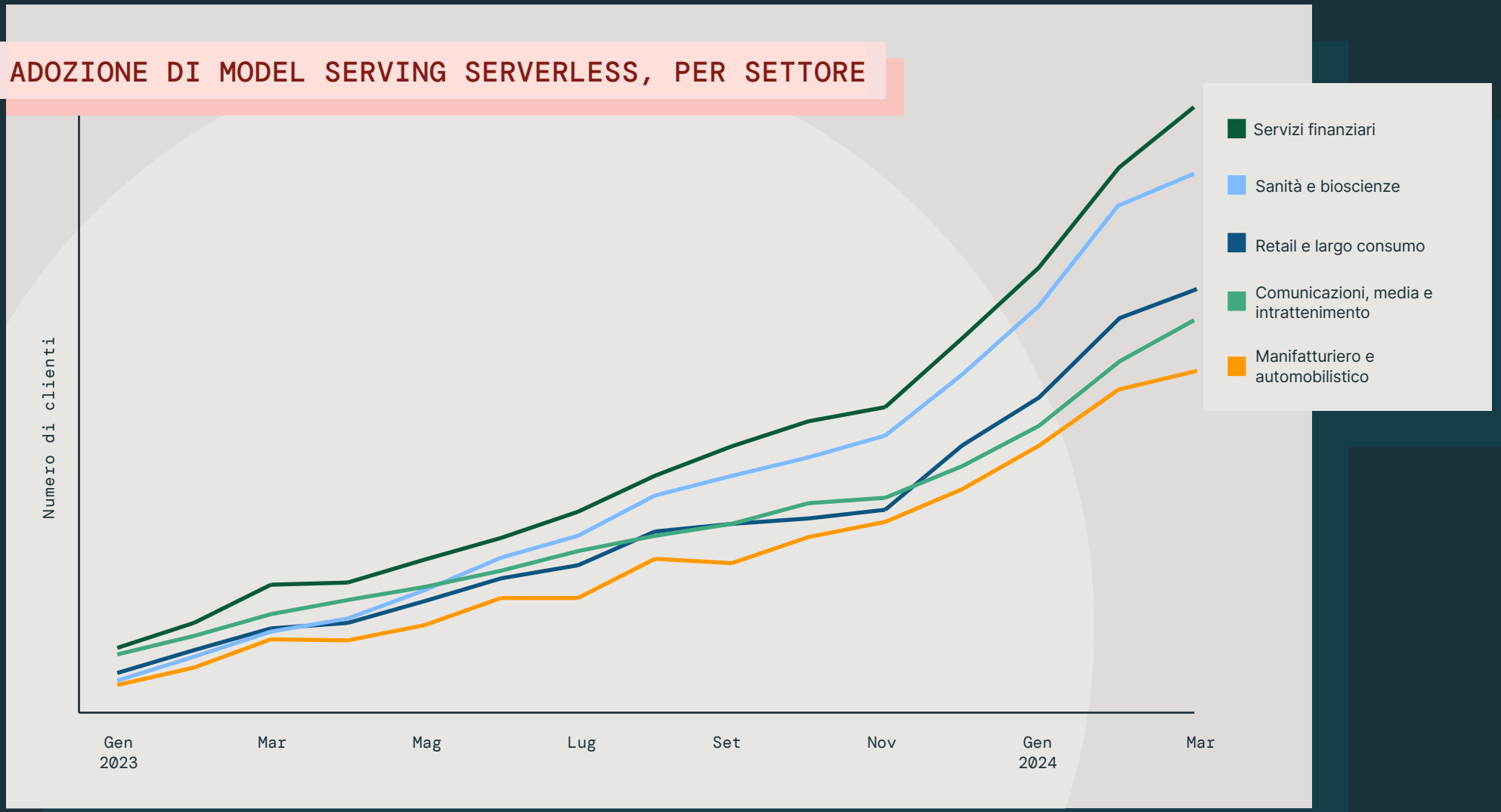


Figura 11: Il settore finanziario primo nell'adozione di prodotti serverless, seguito da sanità e bioscienze.

NOTE: Questi includono model serving su endpoint serverless, Databricks SQL, monitoraggio Lakehouse e job serverless. Nel novembre 2023, l'elaborazione serverless è stata estesa a ulteriori piattaforme cloud regionali.

Le aziende passano al serverless per costruire applicazioni di ML in tempo reale

I sistemi di machine learning in tempo reale stanno rivoluzionando l'operato delle aziende, rendendo possibile fare previsioni o intraprendere azioni immediate sulla base di dati in ingresso. Tuttavia, c'è bisogno di un'infrastruttura di servizio veloce e scalabile la cui costruzione e manutenzione richiede conoscenze specialistiche.

Il model serving serverless scala automaticamente in base alle variazioni della domanda, riducendo così i costi perché le aziende pagano solo per il consumo effettivo. Le aziende possono sviluppare applicazioni di machine learning in tempo reale che vanno dalle raccomandazioni personalizzate al rilevamento delle frodi. Il model serving aiuta inoltre a supportare le applicazioni LLM nelle interazioni con gli utenti.

Abbiamo riscontrato un aumento costante nell'adozione di data warehousing e monitoraggio serverless, scalabili in base alla domanda.

Il settore dei servizi finanziari, il principale utilizzatore di prodotti serverless, ha registrato un aumento di utilizzo del 131% in 6 mesi. Questo settore cerca di anticipare l'andamento dei mercati e le previsioni in tempo reale offrono analisi di mercato più solide.

Il settore sanità e bioscienze ha registrato un aumento nell'utilizzo dei prodotti serverless del 132% in 6 mesi, passando dal quarto al secondo posto nell'ultimo anno. Questo settore deve affrontare notevoli fluttuazioni nei requisiti di elaborazione dei dati, soprattutto durante i periodi di picco o quando si elaborano grandi set di dati, come quelli genetici o di diagnostica per immagini.

Conclusioni

Data science e AI stanno spingendo le aziende verso una maggiore efficienza mentre la GenAI apre un nuovo ventaglio di possibilità. Le piattaforme di data intelligence offrono a tutti i membri di un'organizzazione un luogo coeso e gestito in cui usare dati e AI. I nostri dati mostrano che questi strumenti vengono adottati da aziende di ogni settore e che alcuni dei primi utilizzatori provengono da ambiti del tutto inaspettati.

Le organizzazioni hanno ottenuto guadagni misurabili nella messa in produzione dei modelli di ML. Un numero sempre crescente di aziende adotta e utilizza l'NLP per ricavare informazioni dai dati e si serve di database vettoriali e applicazioni RAG per integrare i dati aziendali nei propri LLM. Gli strumenti open source sono il futuro e continuano a classificarsi ai primissimi posti tra i nostri prodotti più popolari. Le aziende stanno sviluppando strategie basate su una governance unificata di dati e AI.

Considerazioni finali: Le organizzazioni di ogni settore che utilizzeranno in modo più efficace dati e AI otterranno maggiore successo.

Informazioni su Databricks

Databricks è un'azienda di dati e AI. Più di 10.000 organizzazioni in tutto il mondo (fra cui Block, Comcast, Condé Nast, Rivian, Shell e oltre il 60% delle aziende Fortune 500) si affidano alla Databricks Data Intelligence Platform per gestire efficacemente i propri dati e renderli produttivi grazie all'AI. Databricks ha la sede principale a San Francisco e uffici in tutto il mondo ed è stata fondata dai creatori di Lakehouse, Apache Spark™, Delta Lake e MLflow.

Per maggiori informazioni, segui Databricks su LinkedIn, X e Facebook.

databricks