



5 BEST PRACTICE PER LO SVILUPPO DEL DATA WAREHOUSE

EBOOK

SOMMARIO

- 3** Introduzione
- 4** Creare un modello dei dati
- 7** Adottare una metodologia Agile per il data warehouse
- 8** Preferire ELT a ETL
- 9** Adottare uno strumento per l'automazione
- 10** Formare il personale sui nuovi approcci
- 11** Riepilogo
- 12** Informazioni su Snowflake



INTRODUZIONE

La tecnologia cloud ha rivoluzionato il modo in cui le aziende accedono ai dati, li archiviano e li analizzano. Che si tratti di creare una nuova piattaforma dati partendo da zero o di riprogettare un data warehouse legacy per sfruttare nuove funzionalità, alcune linee guida e best practice possono contribuire a garantire il successo del tuo progetto. Alcune di queste best practice possono sembrare ovvie, ma troppo spesso le aziende non dedicano il tempo necessario a stabilire e documentare queste decisioni iniziali, esponendosi così a difficoltà e inefficienze nel futuro.

In questo ebook troverai cinque suggerimenti per strutturare la tua strategia per i dati e allineare l'intera azienda in modo da creare un data warehouse che soddisfi le esigenze aziendali presenti e future. Queste best practice per lo sviluppo del data warehouse aumenteranno le probabilità di creare un data warehouse in grado di fornire più valore a tutti gli stakeholder aziendali, oltre a porre le basi per una più ampia piattaforma dati enterprise in grado di crescere e adattarsi al mutare delle esigenze della tua azienda.

1. CREARE UN MODELLO DEI DATI

Il primo passo per qualsiasi programma incentrato sui dati è creare un modello dei dati: una rappresentazione astratta che organizzi gli elementi dati e ne descriva le relazioni con gli altri elementi e con le proprietà delle entità del mondo reale a cui corrispondono. Un modello dei dati stabilisce una base concettuale condivisa e una definizione chiara delle informazioni che sono importanti per la tua azienda, e del panorama complessivo dei dati aziendali. Un modello dei dati mette a tua disposizione un modo per documentare i data set che saranno incorporati nel data warehouse, le relazioni tra di essi e i requisiti aziendali che la piattaforma dovrà soddisfare.

È possibile creare un data warehouse senza un modello dei dati? Sì, ma tralasciando questo passaggio fondamentale si perdono molti insight preziosi. La creazione di un modello dei dati completo è spesso un'operazione illuminante per le aziende, poiché costringe vari team funzionali a stabilire di comune accordo la definizione e la demarcazione dei data asset e dei requisiti aziendali del data warehouse prima di iniziare lo sviluppo.

Un modello dei dati ben definito continua a produrre risultati positivi molto tempo dopo il lancio in produzione del data warehouse (o data mart). Ad esempio, un modello dei dati stabilisce la derivazione dei dati per tutti gli oggetti contenuti nel data warehouse, facilitando l'inserimento di nuovi membri del team o l'aggiunta di nuovi oggetti dati al data warehouse con il mutare delle esigenze aziendali.

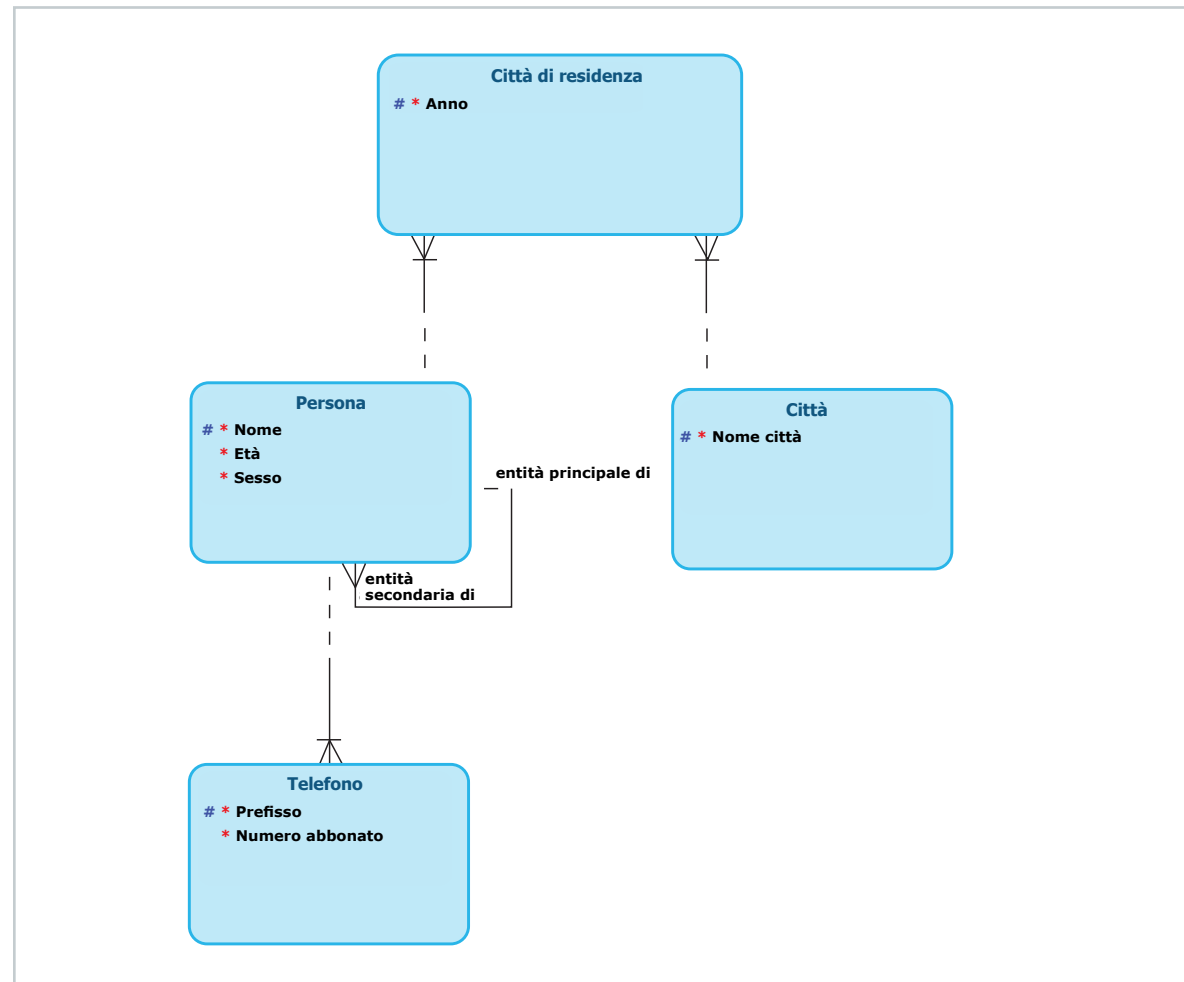


Figura 1: un tipico modello logico dei dati in 3NF (terza forma normale)

Il modello dei dati fornisce inoltre una chiara documentazione del contenuto, del contesto e delle fonti. Questo facilita le verifiche e la conformità a nuovi requisiti relativi ai dati, come quelli previsti dal GDPR, il framework normativo generale per la protezione dei dati dell'Unione europea che stabilisce le linee guida per la raccolta e l'elaborazione dei dati personali delle persone fisiche.

Un modello dei dati robusto contribuirà inoltre a evitare confusione e costose riprogettazioni in tempi successivi. È sempre consigliabile incorporare un livello di integrazione indipendente dalla fonte che consenta di eseguire analisi su più data set in base ai loro elementi comuni.

Un data warehouse riunisce molte fonti e molti tipi di dati diversi, che comprendono data set tradizionali, come quelli relativi alla gestione delle relazioni con i clienti (CRM) e alla pianificazione delle risorse d'impresa (ERP), e data set come blog, feed di Twitter, dati IoT e perfino tipi di dati che non sono ancora stati inventati. Per questo motivo, prevedere un livello di integrazione flessibile, che non dipenda troppo strettamente da un singolo sistema, aiuta a preparare il tuo data warehouse per il futuro.

Un modello dei dati altamente efficiente dovrebbe impiegare definizioni e strutture semantiche determinate dal dominio aziendale, non dalle definizioni specifiche di un singolo sistema di origine. Ad esempio, un sistema CRM potrebbe riferirsi ai clienti con l'etichetta "cli", mentre un altro potrebbe usare "ID_clien". Per il successo del data warehouse è fondamentale stabilire una regola semantica unificata che definisca come gli utenti in tutta l'azienda dovranno designare, accedere e analizzare questi dati in tutti i data set.

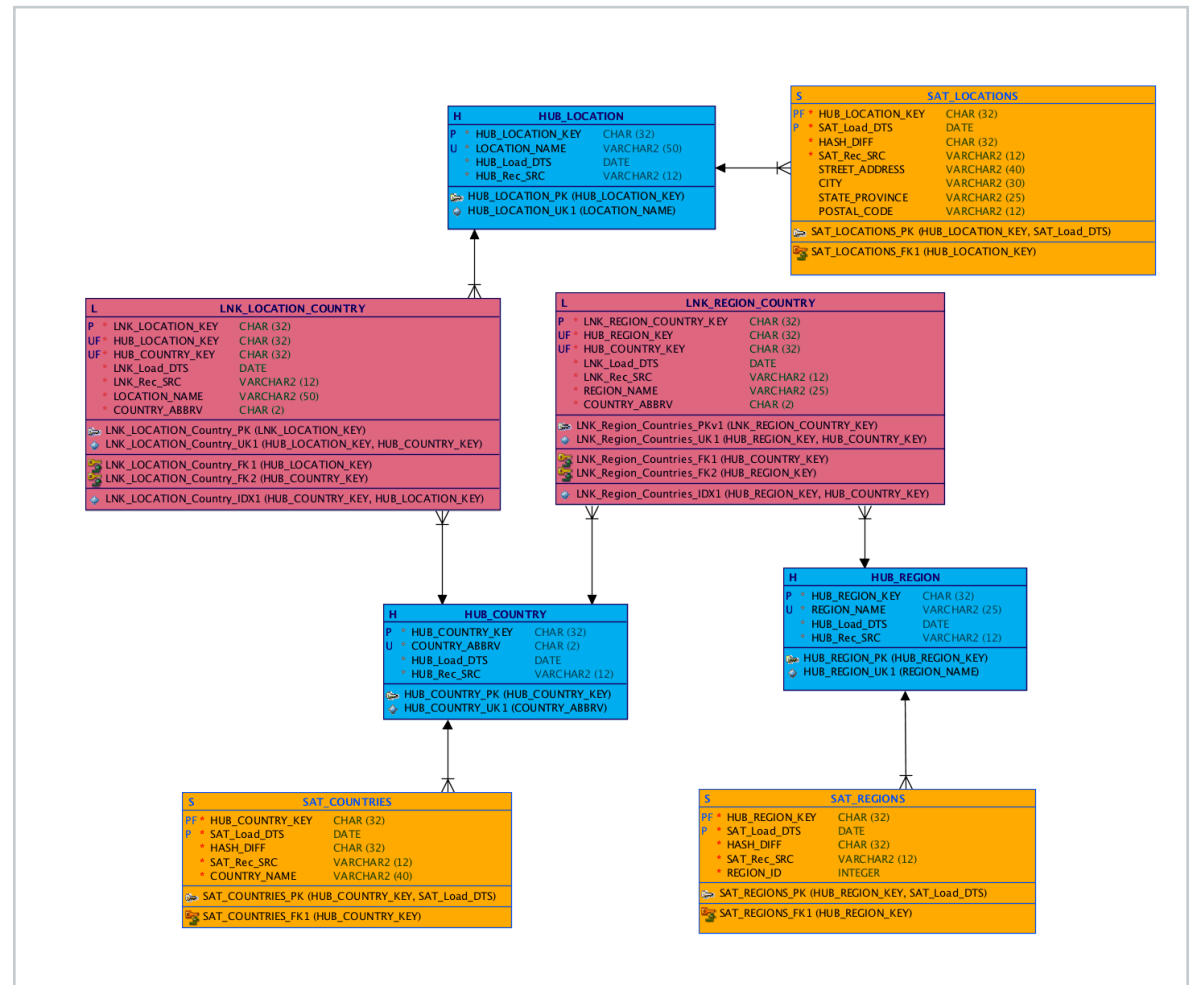


Figura 2: un esempio di modello dei dati che utilizza l'approccio di modellazione Data Vault

Il sistema CRM utilizzato oggi dall'azienda potrebbe essere sostituito da un sistema diverso in seguito a cambiamenti, fusioni o acquisizioni. Se il tuo modello dei dati è rigidamente associato a uno specifico sistema di origine, per integrare il secondo sistema di origine che va a sostituire il sistema legacy sarà necessario un notevole sforzo di riprogettazione. Un livello di integrazione indipendente dalla fonte facilita notevolmente la mappatura dei dati, consentendo di sostituire un vecchio sistema di origine con uno nuovo senza conseguenze per i report a valle e senza dover modificare i comportamenti degli utenti.

All'interno del modello dei dati, è essenziale selezionare un approccio standard. I tipi principali di standard per la modellazione dei dati utilizzati nella progettazione dei data warehouse includono:

- **3NF:** abbreviazione di "3rd Normal Form" (terza forma normale), un'architettura standard progettata per ridurre la duplicazione dei dati e garantire l'integrità referenziale del database.¹
- **Schema a stella:** l'architettura più semplice e più ampiamente utilizzata per sviluppare data warehouse e data mart dimensionali, composta da una o più tabelle di fatti che fanno riferimento a tabelle di dimensioni.²
- **Data Vault (DV):** sviluppata specificamente per risolvere i problemi di agilità, flessibilità e scalabilità riscontrati negli altri approcci, la modellazione DV è stata creata per fornire un repository cronologico granulare, non volatile, verificabile e facilmente estensibile dei dati aziendali. È altamente normalizzata e combina elementi dei modelli 3NF e a stella.³

Ogni architettura offre vantaggi specifici, ma la scelta dell'architettura giusta dipende dalle esigenze specifiche dell'organizzazione.

In realtà, la cosa più importante non è quale architettura verrà adottata, ma il fatto di selezionare, documentare e supportare continuamente questa architettura nel quadro dello sviluppo di un modello dei dati per il data warehouse. Questo consentirà di operare in maniera efficiente in futuro e applicare un'unica metodologia di supporto e risoluzione dei problemi grazie alla quale i nuovi membri del team potranno diventare più rapidamente produttivi.



2. ADOTTARE UNA METODOLOGIA AGILE PER IL DATA WAREHOUSE

In passato, la creazione di un data warehouse (o anche di un data mart) era una vasta operazione monolitica della durata di più trimestri o più anni, soggetta al tradizionale processo "a cascata". Oggi questo modello non è più dominante e molte organizzazioni optano invece per un approccio alla progettazione più flessibile e iterativo, ossia Agile.

Con l'evoluzione sempre più veloce delle esigenze aziendali e l'incalzante comparsa di nuove fonti di dati, le aziende devono essere in grado di adattarsi e sfruttare questi input in modo rapido e conciso. Ciò significa imparare a creare soluzioni per i dati e le analisi in modo incrementale e Agile. Con una pianificazione adeguata e allineata a un unico livello di integrazione indipendente dalle fonti, oggi è possibile suddividere grandi progetti di dati in parti più piccole che possono essere consegnate con maggiore frequenza, per fornire valore aziendale incrementale molto più velocemente.

Per raggiungere questo obiettivo, gli architetti di data warehouse stanno adottando la metodologia Agile, emersa inizialmente nel mondo dello sviluppo software. La metodologia Agile comporta l'evoluzione di requisiti e soluzioni attraverso l'impegno collaborativo di team auto-organizzati e interfunzionali e dei clienti. Applicata all'ideazione e alla realizzazione del data warehouse, questa metodologia consente alle aziende di attivare nuovi data set e risolvere nuove sfide aziendali più rapidamente.⁴

All'interno della metodologia Agile sono emersi numerosi approcci volti a fornire valore più rapidamente, tra cui:

- **Scrum:** il framework di processo più ampiamente utilizzato per lo sviluppo Agile, che prende il nome dal pacchetto di mischia nel rugby in cui i giocatori delle prime linee avanzano insieme a stretto contatto. Questo framework "leggero" mette in primo piano la comunicazione quotidiana e il riesame flessibile dei piani che vengono realizzati nel corso di brevi fasi di lavoro iterative.⁵ Ralph Hughes ha codificato l'applicazione del framework Scrum ai data warehouse in una serie di scritti estremamente influenti, che sono molto utili per le aziende che adottano questo approccio.
- **Kanban:** un metodo per gestire la creazione di prodotti che pone l'accento sulla consegna continua senza sovraccaricare il team di sviluppo. Come Scrum, Kanban è un processo progettato per aiutare i team a collaborare in modo più efficace. Questo metodo prende il nome dalle schede utilizzate per monitorare l'avanzamento della produzione nelle fabbriche giapponesi ed è stato creato da Taiichi Ohno, un ingegnere industriale presso Toyota, per migliorare l'efficienza nel settore manifatturiero.
- **BEAM:** il metodo BEAM (Business Event Analysis and Modelling) è stato introdotto da Lawrence Corr e Jim Stagnitto nel rivoluzionario libro "Agile Data Warehouse Design". Per modellare l'intera area dei processi aziendali, il metodo BEAM si concentra sugli eventi aziendali anziché sui requisiti di reporting noti. Sfrutta sette tipi di dimensioni (chi, cosa, quando, dove, come, quanti e perché) per identificare e quindi elaborare gli eventi aziendali.⁶

Per sfruttare compiutamente i vantaggi dello sviluppo Agile, è molto utile disporre di una piattaforma dati Agile. Le piattaforme dati basate su cloud assicurano questo tipo di elasticità e flessibilità strutturale, consentendo di scalare rapidamente con l'evoluzione delle esigenze aziendali. Le piattaforme dati basate su cloud richiedono meno lavoro, manutenzione e amministrazione per essere utilizzate efficacemente e possono crescere e adattarsi ai requisiti mutevoli dell'azienda. Sfruttando un servizio cloud moderno, i team possono dedicare meno tempo al tuning delle query e al provisioning dello storage e più tempo a risolvere le sfide immediate dell'azienda e fornire valore aziendale.

Sfruttare le metodologie e le strutture agili non è cosa da poco. Richiede un impegno a livello di cultura all'interno dell'organizzazione e spesso rappresenta un cambiamento significativo di mentalità e flussi di lavoro rispetto a quelli tradizionalmente associati al data warehouse. Ristrutturare un team IT perché possa lavorare a pieno ritmo in un ambiente Agile può richiedere da 6 a 12 mesi, una realtà che può sembrare paradossale quando si considera che la metodologia Agile ha l'obiettivo di fornire valore più rapidamente. È possibile accelerare questa transizione avvalendosi della collaborazione di un istruttore esperto di questa metodologia. Una volta completata la transizione, tuttavia, i team potranno iniziare a fornire modifiche incrementali al data warehouse nel giro di settimane anziché mesi.

3. PREFERIRE ELT A ETL

In passato, lo sviluppo dei data warehouse adottava un approccio ETL, ossia "estrai-trasforma-carica": i team estraevano i dati da importare nel data warehouse dai sistemi di origine, li ripulivano o vi applicavano regole aziendali su un server esterno e quindi li caricavano nel data warehouse di destinazione. L'incremento della potenza di elaborazione e delle funzionalità delle piattaforme dati ha portato a preferire il nuovo approccio ELT: "estrai-carica-trasforma".

Nell'approccio ELT, i dati vengono estratti dalla fonte e caricati senza modifiche significative nell'area di staging del data warehouse. I team possono aggiungere metadati, data di caricamento o informazioni sulla fonte a questi dati, che vengono poi trasferiti direttamente nel data warehouse. Una volta caricati i dati nel data warehouse, le aziende possono utilizzare la potenza del database per eseguire trasformazioni, che si tratti di modificare la struttura dei dati (ossia applicare un modello dei dati), applicare regole aziendali o pulire i dati per migliorarne la qualità (ad esempio correggendo indirizzi incompleti, standardizzando i nomi dei campi di dati e risolvendo i duplicati).

L'approccio ELT presenta due chiari vantaggi: risparmio sui costi e maggiore tracciabilità. ELT aiuta a risparmiare sui costi consentendo alle aziende di sfruttare la potenza della piattaforma per trasformare i dati invece di utilizzare un server esterno. La potenza di calcolo basata su cloud è tipicamente molto meno costosa delle operazioni di trasformazione e manipolazione dei dati su un server esterno, quindi il trasferimento diretto dei dati nel cloud risulta più rapido ed economico. L'approccio ELT facilita inoltre le verifiche e la tracciabilità dei dati nel futuro, poiché fornisce un'immagine dei dati originali direttamente all'interno della piattaforma dati. In questo modo, lo stesso data warehouse può ricoprire il ruolo di un "data lake", in cui i dati grezzi vengono conservati in permanenza.



4. ADOTTARE UNO STRUMENTO PER L'AUTOMAZIONE

L'obiettivo del data warehouse è attivare e fornire i dati più rapidamente perché possano informare le decisioni aziendali e generare più valore. Un modo per accelerare la velocità di fornitura è adottare la metodologia Agile. Un altro modo è adottare strumenti di automazione che aiutano a sviluppare e distribuire più rapidamente il codice. Poiché molte metodologie per il data warehouse sono basate su modelli, il codice richiesto per caricare e strutturare i dati è spesso ripetibile e può quindi essere automatizzato. Esiste sul mercato un numero crescente di strumenti in grado di automatizzare alcune o anche tutte le operazioni di progettazione e creazione del codice.

L'automazione consente alle aziende di sfruttare al meglio le proprie risorse, iterare più rapidamente e applicare più facilmente gli standard alla creazione del codice. Consente di creare codice standardizzato, incredibilmente utile per le organizzazioni nelle quali in precedenza il codice e i modelli dei dati ETL venivano sviluppati manualmente. L'automazione fornisce uno standard documentato per questi diversi elementi, oltre a un meccanismo di applicazione e controllo della qualità per garantire che tale standard sia seguito da tutti gli sviluppatori e i progettisti.

Gli strumenti di automazione che utilizzano modelli per generare il codice risultano particolarmente utili, poiché applicano gli standard impostandoli come proprietà preferite all'interno dei modelli stessi. Questo accelera l'onboarding, poiché i nuovi sviluppatori e progettisti utilizzeranno questi strumenti standard, garantendo un'implementazione uniforme e accelerando la curva di apprendimento. Un'implementazione uniforme offre inoltre il vantaggio di semplificare le operazioni di test e debug, poiché il codice viene sviluppato utilizzando gli stessi standard.

Grazie a questi strumenti anche l'iterazione diventa più veloce, poiché i generatori di codice automatizzati normalmente non producono errori di sintassi. L'aggiornamento del codice tipicamente comporta l'aggiunta di un nuovo oggetto allo strumento o la modifica delle proprietà del modello a livello globale, generando nuovo codice che può essere immediatamente distribuito nell'ambiente per essere testato e convalidato.

5. FORMARE IL PERSONALE SUI NUOVI APPROCCI

Il passaggio alla metodologia Agile o allo sviluppo automatizzato del codice non comporta solo un cambiamento di competenze, ma anche un cambiamento di mentalità. Affinché il team possa sfruttare efficacemente questi nuovi approcci e tecnologie, sarà necessario fare formazione. Questo può significare invitare esperti esterni per formare i team sulle best practice di Scrum o istruire i team sui vantaggi, le regole e le best practice per l'architettura standard che l'azienda ha scelto per la sua piattaforma dati.

Sono disponibili molte risorse del settore per aiutare a gestire la transizione alla metodologia Agile. La **Agile Alliance**, un'associazione globale senza scopo di lucro impegnata a promuovere i concetti dello sviluppo di software Agile descritti nell'omonimo Manifesto, offre molte opzioni di formazione per introdurre i fondamenti di questa metodologia. La **Scrum Alliance** offre certificazioni e formazione di livello elementare e avanzato per il framework Scrum. Analogamente, partner selezionati offrono corsi di formazione e certificazioni per Data Vault attraverso la **Data Vault Alliance**.

Come per qualsiasi nuovo processo e mutamento culturale, le organizzazioni dovranno gestire la curva di adozione per garantire un passaggio uniforme ed efficace al nuovo approccio nelle operazioni quotidiane. Identificando progetti pilota o prove di concetto che permettano di familiarizzare con i nuovi approcci, i team svilupperanno e perfezioneranno le competenze necessarie in scenari protetti ma reali, accelerando l'acquisizione e la padronanza di queste nuove capacità.



RIEPILOGO

Tutte le best practice descritte in questo ebook richiedono un investimento iniziale per realizzare il loro valore aziendale potenziale a lungo termine. Questo investimento, tuttavia, produce un duplice ritorno: pone fin da subito le basi per il successo di un programma di analisi dei dati e in seguito accelera la generazione di valore aziendale incrementale nel tuo ambiente dati, anche molto tempo dopo l'iniziale rilascio in produzione.

Con l'evoluzione dei requisiti aziendali e il desiderio di ottenere un maggiore valore da sempre più dati e tipi di dati, queste best practice ti consentiranno di pensare e crescere al di là dei tradizionali casi d'uso del data warehouse. Con una base solida e una piattaforma agile, potrai includere nuovi tipi di dati nel tuo programma di analisi dei dati e ampliarlo per soddisfare nuove esigenze con il supporto di data science, machine learning, AI e magari anche la monetizzazione dei dati. Con le risorse cloud flessibili e scalabili disponibili oggi, non ci sono davvero limiti a ciò che è possibile fare con i tuoi dati.





INFORMAZIONI SU SNOWFLAKE

Snowflake permette a ogni organizzazione di mobilitare i propri dati con Snowflake Data Cloud. I clienti utilizzano il Data Cloud per unificare i dati contenuti nei silos, esplorare e condividere in totale sicurezza i dati ed eseguire diversi workload analitici. Ovunque siano i dati o gli utenti, Snowflake offre un'esperienza sui dati unica che si estende a più cloud e aree geografiche. Migliaia di clienti di ogni settore, tra cui 573 della classifica 2022 Forbes Global 2000 (G2K) al 31 gennaio 2023, utilizzano il Data Cloud di Snowflake per far crescere le loro aziende.

Scopri di più su [snowflake.com](https://www.snowflake.com)



© 2023 Snowflake Inc. Tutti i diritti riservati. Snowflake, il logo Snowflake e tutti gli altri nomi di prodotti, funzioni e servizi Snowflake menzionati nel presente documento sono marchi o marchi registrati di Snowflake Inc. negli Stati Uniti e in altri Paesi. Tutti gli altri nomi di marchi o loghi menzionati o usati nel presente documento sono a puro scopo identificativo e possono essere marchi registrati dei rispettivi proprietari. Snowflake non può essere associato, sponsorizzato o sostenuto da tali proprietari.

RIFERIMENTI

¹ en.wikipedia.org/wiki/Third_normal_form

² it.wikipedia.org/wiki/Schema_a_stella

³ snowflake.com/blog/data-vault-modeling-and-snowflake

⁴ Agiledata.org/essays/dataWarehousingBestPractices.html

⁵ scrum.org/resources/what-is-scrum

⁶ bystembuilders.com/beam