



# COME CREARE LA TUA DATA MESH SU SNOWFLAKE

Negli ultimi anni, l'approccio alla gestione dei dati basato sulla data mesh<sup>1</sup> è sempre più diffuso. Aziende di tutti i settori decidono di gestire i loro dati su una data mesh decentralizzata per migliorare l'agilità dei dati ed evitare i colli di bottiglia organizzativi spesso legati agli approcci centralizzati e monolitici.

Questo documento illustra l'approccio di Snowflake alla data mesh, descrive alcune delle funzionalità di Snowflake che sono più utili per la data mesh e presenta le architetture tipicamente scelte dai nostri clienti per implementare una piattaforma dati self-service che supporti domini distribuiti.

La data mesh è principalmente un approccio organizzativo che definisce le responsabilità e il coordinamento tra team di dominio separati e i relativi prodotti di dati. Tuttavia, per applicare in modo realistico il concetto di data mesh ai domini è necessaria la tecnologia giusta.

**“ La data mesh non è una tecnologia, [...] ma serve la tecnologia giusta per offrire un'ampia gamma di funzionalità ai team dei prodotti di dati. I diversi team di dominio e prodotto di dati non devono costruire una propria piattaforma di dati e analisi ripartendo ogni volta da zero così come dobbiamo semplificare il lavoro per i team dei prodotti di dati. Senza tutto questo, l'empowerment e il decentramento sono impossibili.”**

—OMAR KHAWAJA, Global Head BI, Roche (2022)

La piattaforma dati di Snowflake è adottata con successo da molte aziende che seguono l'approccio data mesh. Non esiste un'unica piattaforma tecnologica in grado di fornire una soluzione end-to-end completa a supporto del concetto di data mesh. Tuttavia, Snowflake fornisce molte delle funzionalità necessarie per una piattaforma dati self-service, rendendo possibile un'architettura domain-driven distribuita e consentendo di implementare i dati come prodotto e una governance computazionale federata.

## L'APPROCCIO DI SNOWFLAKE ALLA DATA MESH

Dopo aver collaborato con numerosi clienti alle loro iniziative di data mesh, Snowflake ha adottato il seguente approccio:

- Riconosciamo che la data mesh è prima di tutto una trasformazione organizzativa, che ha molte implicazioni non tecniche, ma spesso richiede anche modifiche a livello di tecnologie e architettura IT.
- È necessario essere pragmatici. Consigliamo ai nostri clienti di non puntare a creare la rete di dati "perfetta", ma di farsi guidare affrontando i loro punti critici e obiettivi specifici. Ad esempio, lo storage poliglotta e l'accesso multimodale sono concetti utili, ma le aziende devono concentrarsi su quelle che sono le loro reali esigenze per ottenere il massimo impatto.
- Iniziare in piccolo, espandersi in modo incrementale e lungo la curva di maturità della data mesh. Ad esempio, si può iniziare con uno o due domini e prodotti di dati per soddisfare un'esigenza aziendale immediata, quindi sfruttare il successo iniziale per espandere la rete.
- Essere consapevoli dei costi e della complessità. Ad esempio, un approccio che si è dimostrato utile è mantenere il set di strumenti della piattaforma dati self-service ridotto al minimo e il più coerente possibile su tutti i domini, e allo stesso tempo soddisfare tutti i requisiti critici dei domini.
- Definire fin dall'inizio incentivi e criteri di successo, inclusi KPI misurabili per domini, prodotti di dati, piattaforma dati self-service e controlli di governance.
- Non esiste una soluzione di data mesh pronta all'uso. Facciamo tesoro della nostra vasta rete di partner per creare soluzioni congiunte in grado di soddisfare le esigenze dei clienti. Ad esempio, una soluzione di data mesh spesso comprende strumenti per la governance dei dati, l'automazione, DevOps e altre aree, anche se non ne parleremo nel dettaglio in questo documento.

<sup>1</sup> [www.thoughtworks.com/en-us/what-we-do/data-and-ai/data-mesh](http://www.thoughtworks.com/en-us/what-we-do/data-and-ai/data-mesh)

## FUNZIONALITÀ CHIAVE DI SNOWFLAKE

Snowflake offre una serie di funzionalità chiave che hanno aiutato i nostri clienti a creare la piattaforma dati self-service per la loro data mesh.

### Snowflake è molto più di un cloud data warehouse

Snowflake è un provider di servizi cloud integrato che offre un'ampia gamma di funzionalità per data engineering, data lake, data warehouse, data sharing e componenti importanti di un tipico ciclo di vita di machine learning.

In particolare, gli utenti possono creare e automatizzare le pipeline di trasformazione dei dati per convertire vari tipi di dati di input in prodotti di dati governati. Snowflake può operare con la stessa facilità su formati di file comuni nei bucket di storage su cloud e sui flussi di input (ad esempio, da Kafka) o

sulle sue tabelle relazionali. I formati di file supportati includono JSON, XML, Parquet, AVRO, Delta Lake<sup>2</sup>, Apache Iceberg<sup>3</sup> e altri. Inoltre, Snowflake supporta dati non strutturati, come immagini, video o altri formati binari. I dati possono essere manipolati nella piattaforma di Snowflake utilizzando SQL, Python<sup>4</sup>, Scala, Java e Javascript, oppure richiamando funzioni esterne sulla piattaforma cloud più ampia.

Snowflake può fornire o meno tutte le funzionalità richieste dai team di dominio, ma racchiude in *un unico* servizio una notevole gamma di funzionalità che altrimenti richiederebbero l'integrazione di *una raccolta* di servizi cloud, operazione che può essere lunga e complessa e richiede personale altamente qualificato.



FIGURA 1: SNOWFLAKE COME PIATTAFORMA UNICA PER DIVERSI TIPI DI DATI E WORKLOAD

<sup>2</sup> In public preview al momento della pubblicazione, agosto 2022.

<sup>3</sup> In private preview al momento della pubblicazione, agosto 2022.

<sup>4</sup> In public preview al momento della pubblicazione, agosto 2022.

## Snowflake è una piattaforma distribuita, non un sistema monolitico

Snowflake è una piattaforma distribuita ma interconnessa, che evita i silos e consente ai team distribuiti di scambiare dati governati in modo sicuro. Come funziona? Un'azienda può creare uno o più account Snowflake, che possono risiedere nella stessa regione cloud o in regioni e piattaforme cloud diverse (figura 2). Ogni account può contenere più database separati, le cui risorse di calcolo e storage possono essere implementate e scalate indipendentemente, in modo distribuito.

Più team di dominio possono lavorare in modo autonomo, utilizzando potenza di calcolo indipendente in database separati o anche in account separati, pur continuando a usare la piattaforma di Snowflake sottostante per condividere i data asset tra loro. Si noti che il concetto di "database" di Snowflake non si riferisce solo a un tradizionale database relazionale, ma include anche tutte le altre funzionalità di Snowflake, come data engineering, data lake, data warehouse, data sharing e data science. L'utilizzo di cluster di calcolo per combinare ed elaborare i dati provenienti da più database o account è una funzionalità fondamentale della piattaforma di Snowflake.

## Snowflake dispone di funzionalità integrate di data sharing e marketplace

All'interno di Snowflake, i produttori di dati possono condividere dati, servizi dati o applicazioni con altri account pubblicando metadati ("elenchi").

Utilizzando i controlli di discovery degli elenchi, i produttori possono condividere privatamente con altri account o gruppi di account oppure pubblicamente attraverso il Marketplace Snowflake. I produttori di dati possono specificare contratti di servizio (Service Level Agreement, SLA) o obiettivi di servizio (Service Level Objective, SLO) per i dati che condividono, ad esempio la frequenza di aggiornamento, la portata della cronologia, la granularità temporale dei dati e altre proprietà che aiutano a definire i dati come prodotto.

Altri team possono effettuare ricerche per scoprire i data asset pertinenti a loro disposizione e ottenere o richiedere l'accesso. Tali data consumer hanno accesso in tempo reale ai dati condivisi, ma questi rimangono controllati dal produttore, che può personalizzare le politiche di accesso o revocare l'accesso in qualsiasi momento. L'accesso ai dati condivisi non richiede un processo ETL o di trasferimento dei dati da parte del produttore o del data consumer. I produttori possono anche pubblicare e condividere tabelle esterne, che sono "viste" di file archiviati all'esterno di Snowflake e che possono facoltativamente includere i formati Delta Lake e Iceberg. Possono inoltre condividere i dati con terze parti all'esterno dell'azienda, anche se non sono clienti attivi di Snowflake. Ad esempio, un produttore di dati può condividerli esternamente tramite un cosiddetto "reader account" Snowflake e l'intera gamma di API supportate. In alternativa, può esportare periodicamente i dati (partizionati) in un bucket di storage su cloud, utilizzando uno dei formati di file attualmente più diffusi.

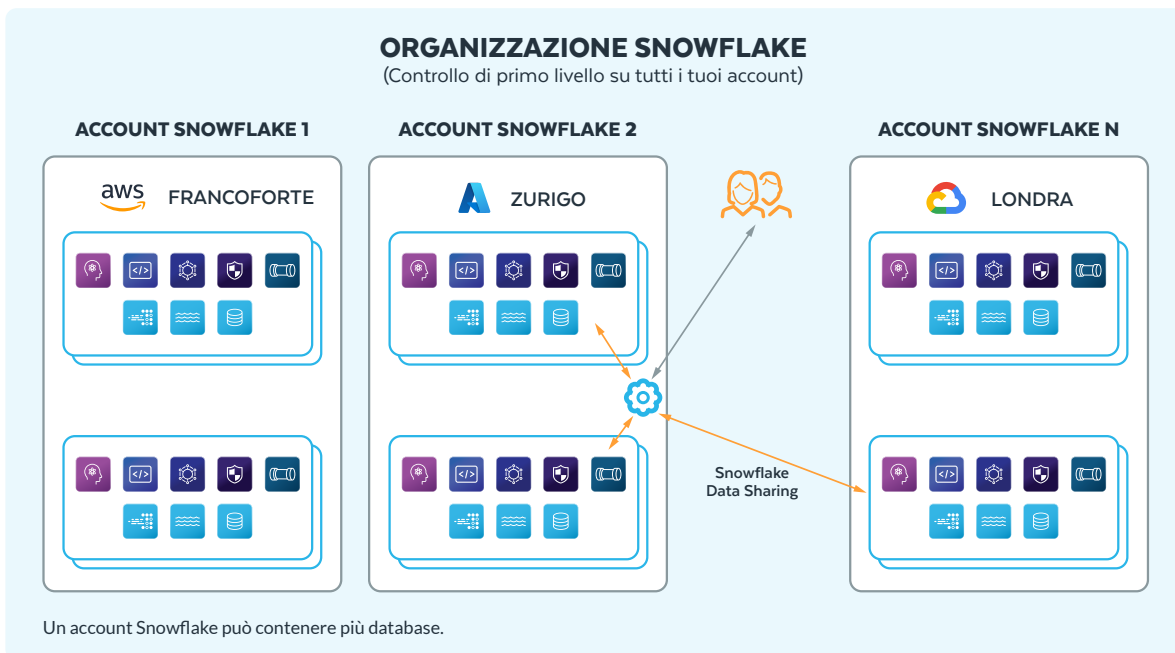


FIGURA 2: L'ORGANIZZAZIONE, GLI ACCOUNT E I DATABASE DI SNOWFLAKE SUPPORTANO UN'ARCHITETTURA DISTRIBUITA

## Snowflake offre un'ampia gamma di funzioni di sicurezza e governance

La governance federata è probabilmente una delle parti più complesse di un percorso verso la data mesh e spesso richiede uno o più strumenti combinati per soddisfare tutti i requisiti. A livello di piattaforma, Snowflake supporta il controllo degli accessi basato sui ruoli, i criteri di accesso a livello di riga, il mascheramento dei dati a livello di colonna, la tokenizzazione esterna, la derivazione dei dati, le funzionalità di verifica e altro ancora. Gli utenti possono anche assegnare uno o più tag di metadati (coppie chiave-valore) alla maggior parte dei tipi di oggetti in Snowflake, come account, database, schemi, tabelle, colonne, cluster di calcolo, utenti, ruoli, attività, condivisioni e altri oggetti. I tag vengono ereditati attraverso la gerarchia degli oggetti e possono essere sfruttati per trovare, tracciare, limitare, monitorare e verificare gli oggetti in base a una semantica definita dall'utente. Inoltre, i criteri di accesso basati su tag<sup>5</sup> consentono agli utenti di associare una limitazione dell'accesso a un tag, in modo che il criterio di accesso venga automaticamente applicato a qualsiasi oggetto di dati corrispondente a cui è assegnato il tag in questione.

In Snowflake, la definizione della maggior parte dei controlli di governance, quali tag, criteri di accesso o regole di mascheramento, può essere definita separatamente dall'applicazione di tali controlli agli oggetti di dati. In questo modo, i proprietari dei domini possono concordare tag o criteri comuni per tutti i domini, che saranno tuttavia applicati o estesi separatamente da ciascun dominio. Inoltre, è possibile utilizzare viste sicure e funzionalità di data clean room per l'analisi dei dati sensibili che altrimenti non potrebbero essere condivisi.

Dal punto di vista dell'utilizzo del prodotto, vengono raccolte metriche come la telemetria e i dati di consumo che possono essere utilizzate per l'analisi dell'impatto. In questo modo, i team di dominio possono monitorare come e con quale frequenza i loro prodotti di dati vengono utilizzati da consumer diversi.

## Snowflake offre un'esperienza quasi self-service

La facilità d'uso e la manutenzione quasi inesistente sono due dei motivi più comuni per cui i clienti scelgono Snowflake. Queste sono proprietà fondamentali per una piattaforma self-service. Ad esempio, gli utenti possono creare facilmente istanze dei propri cluster di calcolo e scalarle senza l'assistenza di un team dell'infrastruttura IT. Clonare gli ambienti di sviluppo e test è altrettanto semplice. È possibile impostare un meccanismo di acquisizione dei dati sulle modifiche con un'istruzione DDL SQL di una sola riga. Questa attenzione alla facilità d'uso è stata un principio guida per lo sviluppo di tutte le caratteristiche e le funzioni della piattaforma di Snowflake.

<sup>5</sup> In public preview al momento della pubblicazione, agosto 2022.

## PRODOTTI DI DATI IN SNOWFLAKE

In una data mesh, ogni dominio crea, gestisce ed è proprietario di uno o più prodotti di dati condivisi con altri domini e data consumer. Per trattare i dati come un prodotto è necessaria soprattutto una mentalità orientata al prodotto, che deve diventare un'abitudine organizzativa. Inoltre, i domini richiedono strumenti self-service adeguati che supportino la creazione e la gestione dei prodotti di dati. Vediamo come Snowflake può aiutarti a implementare il concetto di dati come prodotto.

Si definisce "prodotto di dati" la combinazione dei dati con metadati, codice e dipendenze dell'infrastruttura.

- **Dati:** in Snowflake, i dati di un prodotto di dati possono assumere varie forme, come tabelle, viste, file (JSON, XML, Parquet, Avro, CSV, ecc.) o tabelle esterne che fungono da viste di file esterni a Snowflake. Un singolo prodotto di dati può essere costituito da più oggetti di questo tipo. Una pratica tipica per i domini è utilizzare uno schema per ogni prodotto di dati, per raggruppare gli oggetti di dati e facoltativamente anche il codice di ciascun prodotto. I produttori di dati possono modellare i dati nel modo più adatto a soddisfare le esigenze dei data consumer.
- **Metadati:** i metadati di un prodotto di dati includono i metadati tecnici dei suoi oggetti di dati, come i nomi delle tabelle, i nomi delle colonne, i tipi di dati o le definizioni dei formati di file. Includono inoltre le dipendenze degli oggetti, la derivazione dei dati e la cronologia degli accessi. Ogni oggetto può anche essere contrassegnato da tag, ossia coppie chiave-valore che esprimono metadati arbitrari come origine dei dati, dominio, sensibilità, termini aziendali, tassonomia, centri di costo o altri attributi definiti dall'utente.

Quando un prodotto di dati viene pubblicato nel Marketplace Snowflake, il produttore deve fornire una documentazione che può comprendere descrizione del prodotto, esigenza aziendale, esempi, termini di servizio e un link al supporto per il prodotto di dati. Al produttore viene inoltre richiesto di specificare gli SLO del prodotto di dati, come la frequenza di aggiornamento, la cronologia, la granularità temporale dei dati e altre proprietà (figura 3).

- **Codice:** il codice di un prodotto di dati include le pipeline e le trasformazioni che creano e aggiornano un prodotto di dati. In Snowflake, tale codice può includere oggetti di Snowflake quali task, pipe, stream, stored procedure<sup>6</sup>, funzioni definite dall'utente e così via, che possono essere raggruppati in uno schema per ogni prodotto di dati. Il codice in questi oggetti può essere SQL, Java, Javascript, Scala o Python e viene eseguito in modo nativo nella piattaforma di Snowflake.

Il codice può anche includere criteri, ad esempio il controllo dell'accesso basato sui ruoli, criteri di mascheramento dinamico dei dati, criteri di controllo dell'accesso a livello di riga, viste protette, assegnazione di tag agli oggetti o codice per classificare o anonimizzare/tokenizzare i dati.

- **Dipendenze dell'infrastruttura:** a titolo di esempio, un task Snowflake che pianifica e orchestra la pipeline per aggiornare un prodotto di dati può specificare un determinato cluster di calcolo per il processo. Può trattarsi di una risorsa di calcolo dedicata per un solo prodotto di dati o condivisa tra più prodotti di dati. In entrambi i casi, il cluster può essere sospeso e ripristinato automaticamente secondo necessità, in modo da incorrere in un costo solo quando esegue operazioni. Inoltre, i cluster possono essere scalati verso l'alto e verso il basso in modalità self-service. Task, pipe e altre operazioni possono anche essere serverless per ridurre o eliminare la necessità di dipendenze dell'infrastruttura esplicite.

FIGURA 3: DEFINIZIONE DEGLI SLO PER UN PRODOTTO DI DATI

<sup>6</sup> In public preview al momento della pubblicazione, agosto 2022.

Snowflake supporta una varietà di porte di ingresso e uscita per i prodotti di dati, tra cui l'inserimento in streaming, un connettore Kafka, un connettore Spark, un'API Dataframe, l'inserimento automatico dei dati da bucket di storage su cloud, un'API REST, formati di file e, naturalmente, API SQL come JDBC, ODBC, .NET e API per molti dei linguaggi di programmazione più diffusi. È anche possibile utilizzare le funzionalità di collaborazione di Snowflake per distribuire in modo

sicuro e trasparente dati, servizi dati e applicazioni tra cloud diversi senza richiedere ulteriori integrazioni o pipeline ETL.

I prodotti di dati dovrebbero inoltre avere una serie di importanti proprietà. La tabella 1 elenca alcuni esempi di funzionalità di Snowflake che possono aiutarti a realizzarle.

CARATTERISTICHE DEL PRODOTTO DI DATI	ESEMPI DI FUNZIONALITÀ DI SNOWFLAKE (ELENCO NON COMPLETO)
<b>Sicuro</b>	Controllo degli accessi basato sui ruoli, criteri di accesso a livello di riga, mascheramento dinamico dei dati, crittografia, tokenizzazione
<b>Individuabile</b>	Discovery mirata/Marketplace Snowflake, integrazione opzionale con un catalogo di terze parti
<b>Indirizzabile</b>	Condivisioni di dati Snowflake, accesso standardizzato tra più cloud e regioni diverse
<b>Comprensibile</b>	Tag di metadati personalizzati, prodotti di dati con documentazione, forma statistica dei dati in Snowsight
<b>Affidabile</b>	SLO/SLA per frequenza o granularità degli aggiornamenti, derivazione dei dati, dipendenze degli oggetti, cronologia degli accessi, ecc.
<b>Accessibile in modo nativo</b>	SQL, Python, Java, Scala, API SQL, API REST, DataFrame ecc. per accedere a dati multimodello (strutturati, semi-strutturati, non strutturati, vari tipi di file, ecc.)
<b>Interoperabile</b>	Tipi di dati SQL ANSI, metadati unificati e API comuni tra domini, Snowflake Collaboration, Data Sharing, Marketplace, Data Exchange
<b>Valore indipendente</b>	Prodotti di dati composti formati da più oggetti; possono comprendere oggetti di dati e funzioni condivisibili con gli utenti dei prodotti di dati

TABELLA 1: CARATTERISTICHE DEL PRODOTTO DI DATI SUPPORTATE IN SNOWFLAKE

## OPZIONI DI ARCHITETTURA PER I DOMINI DISTRIBUITI

Esaminiamo ora alcune delle topologie di Snowflake che le aziende hanno scelto come piattaforma per supportare i domini distribuiti. Queste topologie sono quadri generali e le implementazioni effettive possono variare in base ai requisiti e alle preferenze specifiche.

- "Account per dominio": ogni dominio utilizza un account Snowflake separato
  - Massimo isolamento tra i domini.
  - Domini diversi possono operare in regioni e piattaforme cloud differenti.
  - Consente di realizzare una data mesh multi-regione e multi-cloud con un'esperienza Snowflake coerente e funzionalità di condivisione dei dati tra domini integrate utilizzando un Data Exchange centralizzato per i metadati in cui tutti i domini possono pubblicare e ottenere prodotti di dati.
- "Database per dominio": ogni dominio utilizza uno o più database Snowflake separati
  - Tutti questi database sono gestiti in un unico account Snowflake.
  - Gestione semplificata di utenti, sicurezza e governance tra domini diversi.
  - L'accesso ai prodotti di dati può essere fornito facilmente impostando autorizzazioni a livello di oggetto tra database diversi.
  - Ogni team di dominio può comunque creare e scalare i propri cluster di calcolo indipendentemente dagli altri domini.
- "Schema per dominio": ogni dominio utilizza schemi separati in un singolo database
  - Livello minimo di isolamento tra gli ambienti di dominio.
  - Ogni team di dominio può comunque creare e scalare i propri cluster di calcolo, isolati dagli altri domini.
  - Richiede potenzialmente maggiore impegno per definire le convenzioni di denominazione per distinguere tra oggetti provenienti da domini diversi.
  - Utile per i sottodomini in uno scenario che prevede domini e sottodomini.

- Nel seguito di questo documento l'opzione "schema per dominio" non verrà discussa nei dettagli, ma presenta molte somiglianze con l'opzione "database per dominio".

- "Domini eterogenei": i domini possono utilizzare stack IT diversi
  - Alcuni domini utilizzano Snowflake e altri utilizzano altri sistemi.
  - Alcuni domini sono nel cloud e altri possono essere on-premise.
  - In genere comporta un grado di complessità maggiore per adattarsi ad ambienti di dominio eterogenei.
  - Richiede un'attenzione particolare, poiché può essere contrario all'obiettivo della data mesh di utilizzare in tutti i domini una piattaforma self-service comune indipendente dal dominio.

Ulteriori variazioni dell'architettura o approcci ibridi derivati da questi tipi di base sono possibili e praticabili. Ad esempio, un'azienda potrebbe scegliere l'approccio "database per dominio" e gestire questi database in più account anziché in un unico account. Inoltre, alcuni domini potrebbero utilizzare database separati mentre altri utilizzano un intero account Snowflake. In più, l'ambiente utilizzato da un team di dominio è spesso costituito da Snowflake e da strumenti aggiuntivi che variano in base a esigenze e competenze del team.

Il concetto chiave è che Snowflake supporta più architetture che consentono di bilanciare in modi diversi l'autonomia dei domini e il decentramento da un lato e vari gradi di complessità e gestione operativa dall'altro.

Ogni azienda deve trovare l'equilibrio tra centralizzazione e decentralizzazione più adatto alle proprie dimensioni, alla propria infrastruttura esistente e alla propria cultura organizzativa. Lo stesso vale per la governance federata, in cui le aziende devono scegliere l'equilibrio più consono alle proprie esigenze tra controllo centralizzato e autonomia dei domini locali.

Nelle sezioni seguenti vengono descritte in dettaglio queste opzioni di architettura. Questa discussione è incentrata principalmente sulle topologie di Snowflake piuttosto che sull'integrazione con strumenti di terze parti che spesso i nostri clienti utilizzano insieme a Snowflake per le loro iniziative di data mesh.

### Account singolo: database per dominio

Una topologia diffusa, adottata da molti nostri clienti della data mesh, si basa su un unico account Snowflake in cui i vari domini utilizzano database separati e cluster di calcolo separati come ambienti autonomi. A ciascun dominio possono essere assegnati uno o più database e cluster per le sue esigenze di sviluppo, test e produzione. La natura self-service della piattaforma consente ai domini di utilizzare le funzionalità di clonazione zero-copy di Snowflake per (ri)creare ambienti di sviluppo e test in modo istantaneo e frequente. Inoltre, è possibile consentire a utenti diversi all'interno di un dominio di avviare e scalare in modalità self-service i propri cluster di calcolo in base alle rispettive esigenze. È comunque possibile configurare il monitoraggio dei costi e dei consumi e le quote per i domini o per altri livelli di granularità nella gerarchia di utenti e risorse.

Utilizzando il Data Cloud di Snowflake, tutti i domini possono disporre di ambienti e risorse di calcolo separati senza che questi costituiscano silos fisici che renderebbero difficile l'accesso ai prodotti di dati.

Un certo grado di governance viene deciso a livello centrale e applicato a tutti i database con un processo DevOps. Ciò può essere facilitato da funzioni come l'assegnazione di tag agli oggetti per creare in modo semplice una panoramica dei diversi oggetti che appartengono ai domini. All'interno dei domini, la governance è controllata dai team di dominio, che applicano il controllo dell'accesso basato sui ruoli e i criteri di accesso a livello di riga e colonna per proteggere i dati ed escludere utenti e domini dall'accesso indesiderato a determinati dati.

## CATALOGO DI TERZE PARTI / MARKETPLACE SNOWFLAKE

Inventario dei prodotti di dati condivisi

METADATI

RICERCA

FLUSSO DELLA RICHIESTA DI ACCESSO

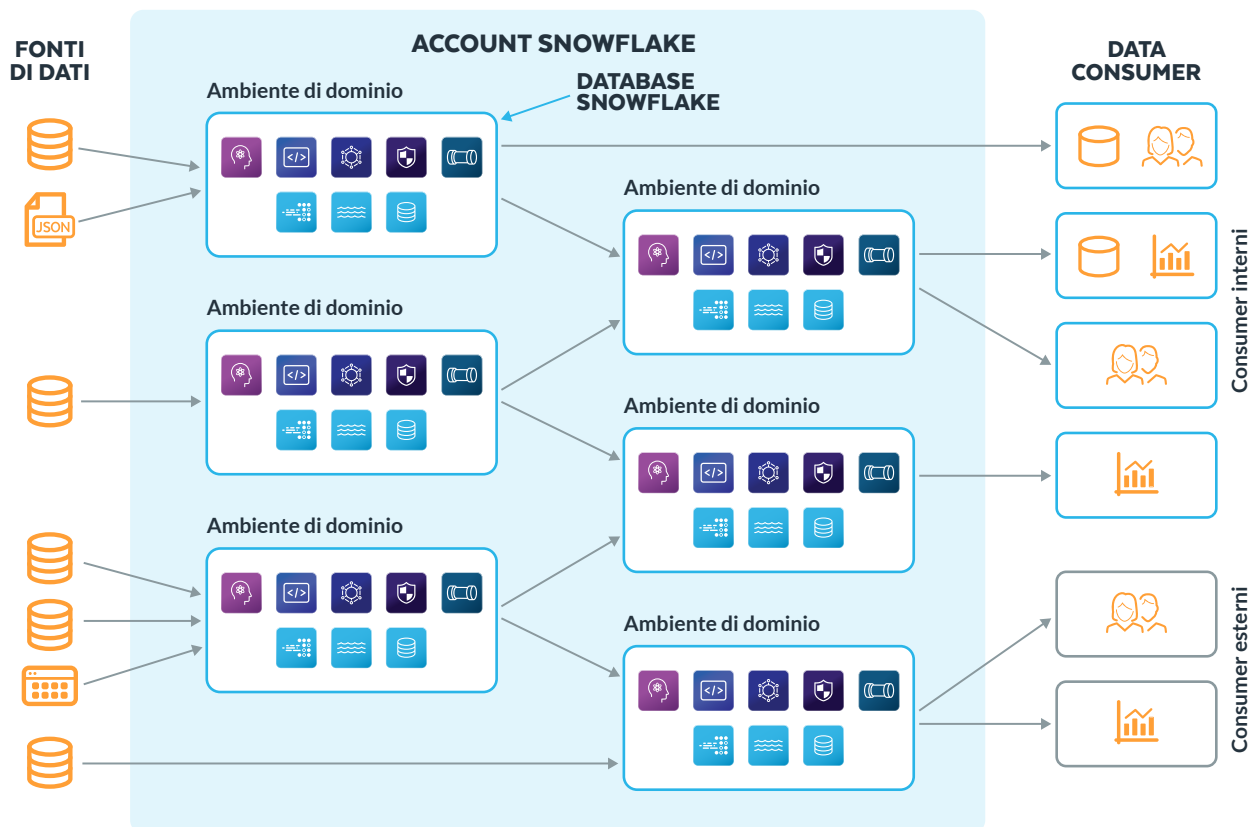


FIGURA 4: ACCOUNT SNOWFLAKE SINGOLO - DATABASE PER DOMINIO

Ogni dominio può avere più schemi, uno dei quali funge da livello in cui rendere disponibili i prodotti ad altri domini. Un altro approccio consiste nell'utilizzare un database condiviso comune, in cui ogni dominio dispone di uno schema per pubblicare i propri prodotti di dati sotto forma di viste (senza creare copie). Questi prodotti possono comprendere dati strutturati, semi-strutturati o non strutturati, a seconda del contenuto del prodotto. I prodotti vengono quindi elencati in un catalogo di dati di terze parti, dove potranno essere individuati.

Abbiamo visto applicare vari metodi per gestire le richieste di accesso a un prodotto, ad esempio un approccio manuale in cui il richiedente deve aprire un ticket che verrà elaborato dal team di dominio, che a sua volta concederà o negherà l'accesso mediante l'assegnazione di un ruolo di accesso al richiedente. Alcuni cataloghi forniscono un flusso più automatico.

L'uso di un unico account Snowflake per tutti i domini offre i seguenti vantaggi:

- L'accesso ai prodotti di dati può essere ottenuto facilmente impostando le autorizzazioni tra database diversi.
- L'amministrazione centralizzata dei criteri di rete, sicurezza e governance semplifica la gestione complessiva.
- Il disaster recovery è più semplice, in quanto richiede solo il supporto di un altro account in un'altra regione o cloud.

Le convenzioni di denominazione devono essere pianificate con attenzione, poiché il numero di oggetti può essere considerevole, dato che ogni dominio può creare facilmente con la clonazione zero-copy tutti gli ambienti di sviluppo, test, accettazione e produzione (Development, Test, Acceptance, Production, DTAP) richiesti.

Un approccio analogo è rappresentato dall'uso di uno schema per dominio. Le implicazioni di questo approccio sono simili all'uso di un database per dominio, in quanto tutto è organizzato logicamente all'interno di un singolo account Snowflake. Si noti, tuttavia, che l'uso di un database per dominio rende più semplice la condivisione dei dati con utenti esterni, sia pubblicamente tramite il Marketplace Snowflake, sia privatamente utilizzando i controlli di discovery degli elenchi di dati.

### Più account: un account per dominio

Un'altra possibile topologia consente a ciascun dominio di operare in un account Snowflake separato. Questi account possono trovarsi nella stessa regione o in regioni e piattaforme cloud diverse. Il Data Cloud globale di Snowflake consente alle aziende e ai domini di condividere i dati tra account, regioni e piattaforme cloud diverse e di ottenere facilmente e in modo sicuro e governato l'accesso standardizzato ai prodotti di dati delle controparti. Alcuni dei nostri clienti utilizzano questa funzionalità per supportare una data mesh multi-regione e multi-cloud.

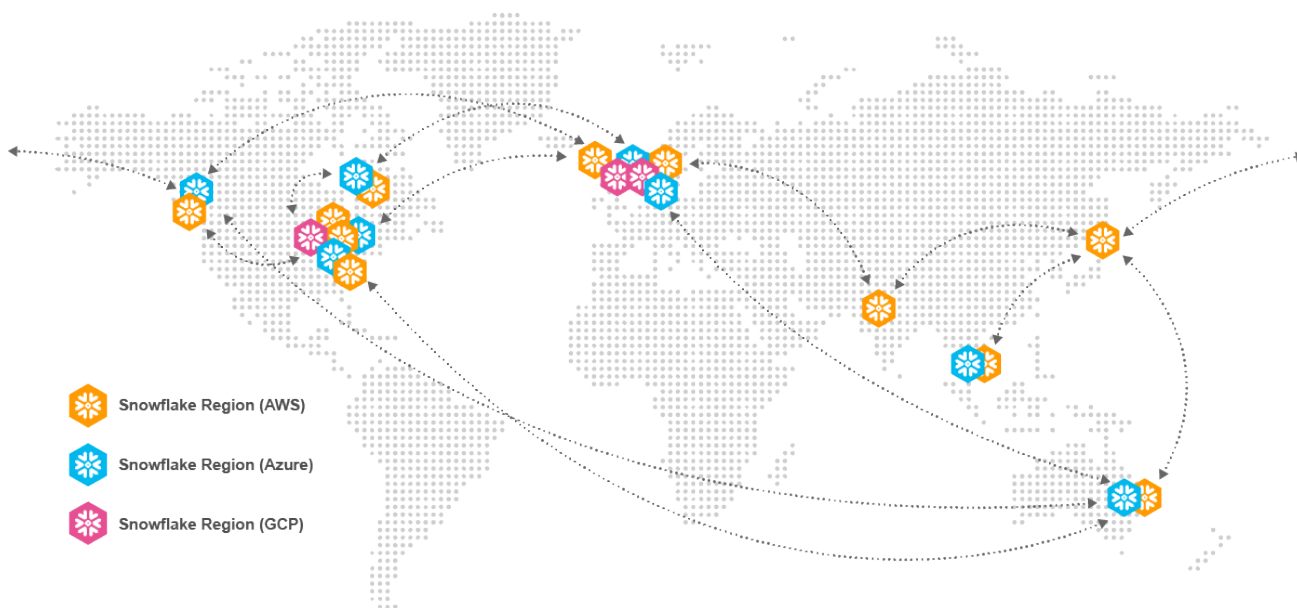


FIGURA 5: SNOWGRID COME PIATTAFORMA DATA CLOUD GLOBALE

Questa topologia viene scelta per diversi motivi. Ad esempio, se un'azienda ha divisioni distribuite a livello globale, i diversi domini potrebbero allinearsi in modo naturale a località e aree geografiche diverse. Alcune aziende operano a livello internazionale e hanno bisogno di rispettare i requisiti di residenza dei dati, ad esempio alcuni dati non possono uscire dall'Europa senza essere sottoposti ad anonimizzazione, mascheramento o altre misure per garantire la conformità alle normative sulla privacy dei dati.

Fusioni e acquisizioni sono altri motivi per cui le aziende si trovano a dover scambiare dati tra regioni o piattaforme cloud diverse. Alcune aziende perseguono intenzionalmente una strategia multi-cloud per diversificare o per tener conto delle preferenze e degli investimenti esistenti di diverse business unit. L'utilizzo di account separati consente inoltre di ottenere una maggiore autonomia dei domini, ad esempio nei casi in cui la gestione degli utenti e della sicurezza devono essere separate per ciascun dominio.

La topologia risultante (figura 6) è logicamente molto simile all'uso di un database separato per ogni dominio, salvo che ogni dominio ora "possiede" un account Snowflake separato e utilizza le funzionalità di condivisione dei dati di Snowflake e il Marketplace Snowflake per rendere i prodotti di dati accessibili agli altri.

I vantaggi rispetto all'uso di un database per dominio possono essere i seguenti:

- Consente di utilizzare le funzionalità di data sharing e collaborazione tra domini diversi.
- È più facile applicare standard di denominazione globali, poiché ogni account ha uno spazio dei nomi indipendente.
- È possibile supportare le preferenze in fatto di regione e piattaforma cloud.
- La gestione della sicurezza e degli utenti è separata per ciascun account.

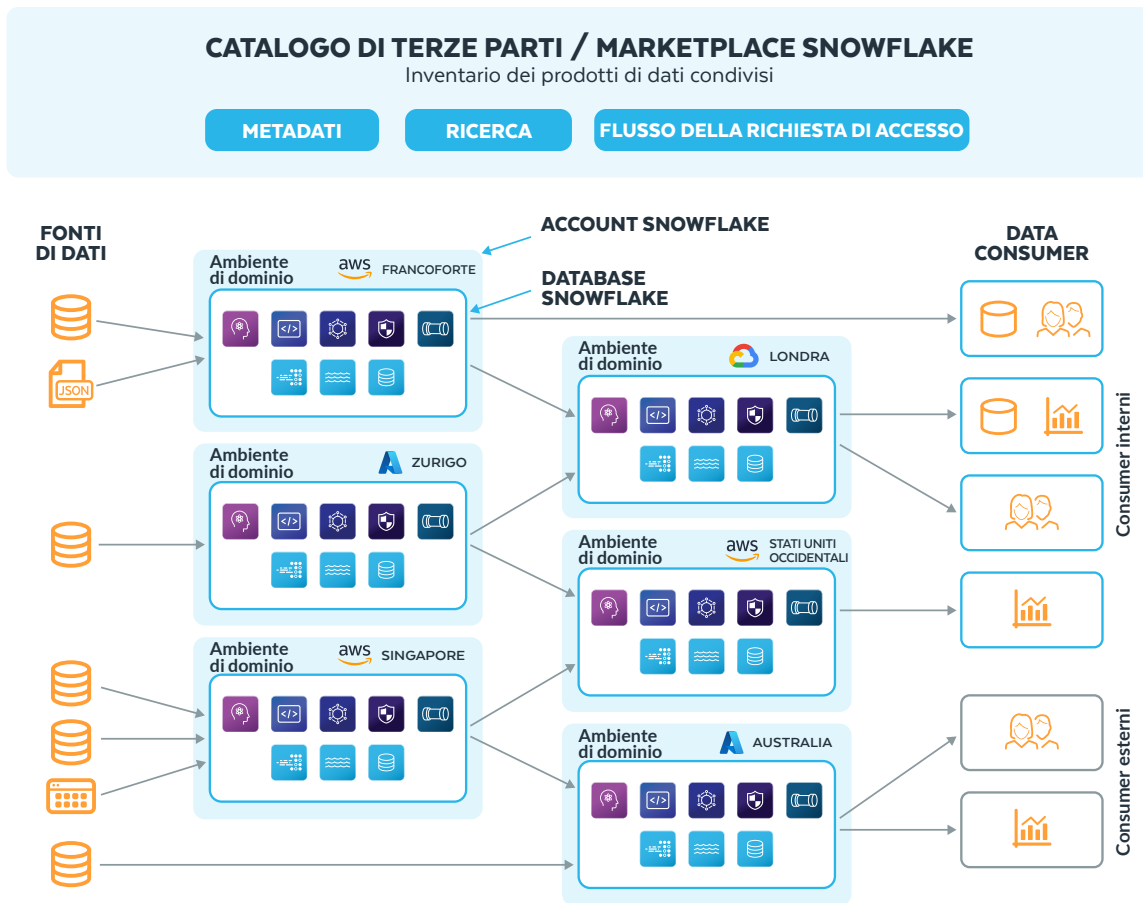


FIGURA 6: PIÙ ACCOUNT - UN ACCOUNT PER DOMINIO

## Architettura eterogenea

Alcuni clienti ci hanno chiesto come integrare altri ambienti di dominio non Snowflake nelle topologie discusse sopra. Tali integrazioni si traducono in un'architettura eterogenea, in cui non tutti i domini utilizzano la stessa piattaforma dati indipendente dal dominio per implementare le pipeline e i prodotti di dati. Spesso questa scelta è motivata dal desiderio di riutilizzare più repository o stack tecnologici già esistenti in parti diverse dell'organizzazione.

In base alla nostra esperienza, in genere un'architettura così eterogenea aumenta i costi e la complessità di un percorso verso la data mesh. Il motivo è che più i sistemi partecipanti sono eterogenei, più è difficile garantire la coerenza in termini di governance, sicurezza, metadati, standard di interoperabilità, prestazioni, competenze richieste, supporto IT e altre aree critiche. Pertanto, incoraggiamo i clienti a considerare attentamente il ruolo dei diversi sistemi e repository che vogliono integrare nella loro data mesh. Questi sistemi sono realmente utilizzati come ambienti di dominio che creano e distribuiscono prodotti di dati, oppure

dovrebbero essere visti come fonti di dati che i domini ricevono come input? Nel secondo caso, spesso i clienti possono scegliere tra le topologie discusse sopra.

Un possibile approccio all'integrazione delle implementazioni di domini non Snowflake consiste nel fare in modo che i data asset o i "prodotti di dati quasi pronti" di tali domini vengano inviati a un livello intermedio, per il quale Snowflake può fungere da "proxy" che distribuisce i prodotti di dati al resto della data mesh garantendo la coerenza della governance, della sicurezza, dell'interoperabilità e così via.

Questo livello intermedio potrebbe essere costituito, ad esempio, da argomenti Kafka che vengono trasmessi a Snowflake attraverso l'ingestion continua seguita da aggiornamenti automatici dei prodotti di dati in Snowflake. Il livello intermedio potrebbe consistere anche in uno o più bucket di storage su cloud in Amazon S3, Azure Blob Storage, Azure Data Lake Storage o Google Cloud Storage. I formati di dati possono includere JSON, XML, Parquet, AVRO, Apache Iceberg, Delta Lake e altri. Snowflake può ricevere automaticamente i nuovi file dai bucket

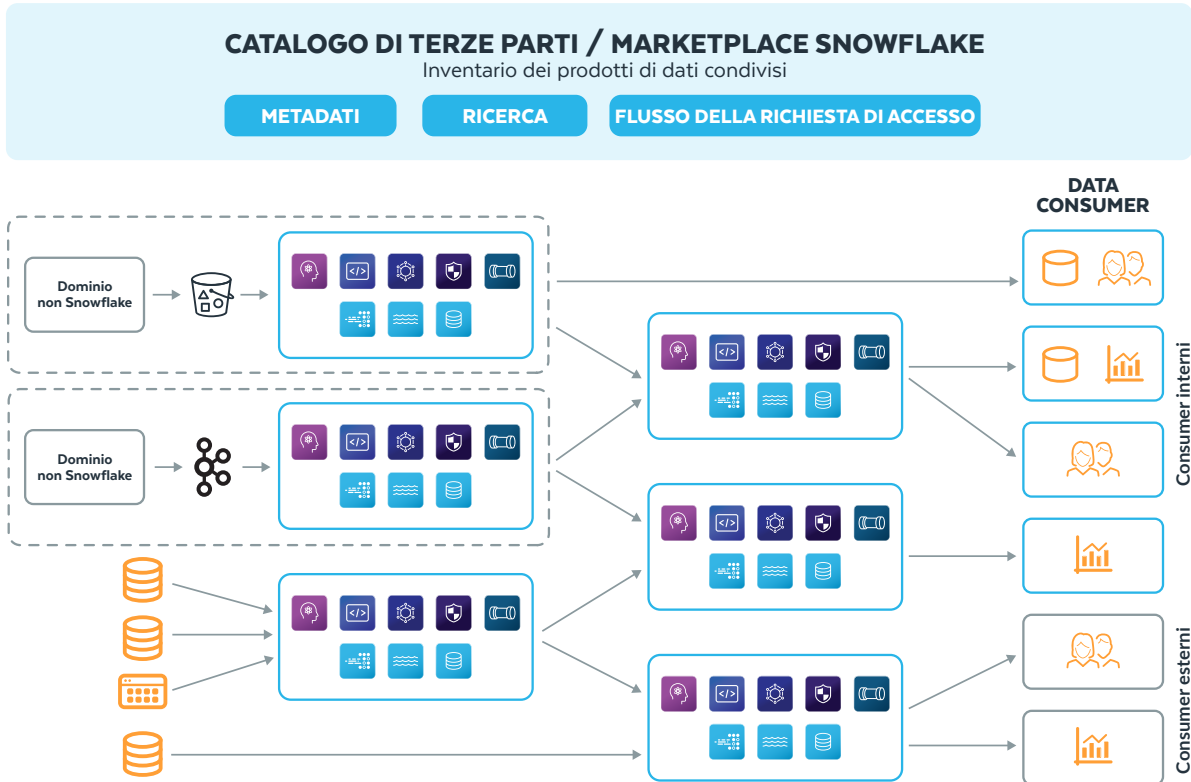


FIGURA 7: ARCHITETTURA ETEROGENEA

di storage in maniera continua per ottimizzare prestazioni, sicurezza e gestione automatica, oppure esporre file come le tabelle esterne al resto della data mesh per l'accesso in lettura. Le tabelle esterne di Snowflake sono essenzialmente viste di dati che risiedono in file esterni a Snowflake, ma sono anche oggetti di dati di prima classe all'interno di Snowflake, che possono essere protetti e governati, uniti e persino condivisi tramite la collaboration di Snowflake, in modo analogo agli altri oggetti di dati in Snowflake. Snowflake funge così da livello di integrazione in grado di esporre i dati esterni in modo coerente e governato senza bisogno di ingestione e copie dei dati.

Esistono anche altre opzioni per l'integrazione di ambienti non Snowflake nella piattaforma dati, che non rientrano nell'ambito di questo documento.

Alcune aziende considerano la virtualizzazione dei dati come una potenziale soluzione per integrare un insieme diversificato di ambienti di dominio. Sebbene esistano casi d'uso validi per la virtualizzazione dei

dati, abbiamo notato che questa scelta comporta anche una serie di sfide, per esempio le prestazioni quando si devono unire i dati di più repository diversi. Questa operazione richiede in genere lo spostamento dei dati per portarli in una posizione comune in cui calcolare l'unione, anche se è possibile inviare altri predicati alle fonti di dati.

Ciò può rendere la virtualizzazione impraticabile per i casi d'uso sensibili alle prestazioni. Una proprietà non trascurabile dell'architettura della piattaforma di Snowflake è che le prestazioni dell'unione di più oggetti di dati in un singolo database sono approssimativamente equivalenti a quelle degli stessi oggetti di dati in database separati o persino in account Snowflake separati. Un'altra sfida della virtualizzazione che abbiamo rilevato in alcune aziende è che può incoraggiare i team a rimanere chiusi nelle rispettive "isole tecnologiche", spesso molto specifiche per il singolo dominio, invece di sforzarsi di creare una piattaforma self-service comune e indipendente dal dominio.

## RIEPILOGO

La data mesh non è una soluzione universale per tutte le sfide legate alla gestione e all'integrazione dei dati. Tuttavia, se stabilisci che la data mesh è l'approccio giusto per la tua azienda, assicurati di concentrarti sugli aspetti organizzativi e non tecnici che sono indispensabili per il successo, come ad esempio cambiamenti organizzativi, ruoli e responsabilità, personale, incentivi e responsabilità, sostegno dei principali stakeholder o riorientamento della mentalità verso il prodotto.

Infine, dovrai progettare un'architettura IT self-service in grado di supportare domini distribuiti e prodotti di dati con governance federata. Self-service e facile da usare, la piattaforma Snowflake può svolgere questo ruolo chiave. Snowflake supporta topologie diverse che consentono alle aziende di scegliere il grado desiderato di decentralizzazione e autonomia dei domini, mantenendoli al tempo stesso interconnessi e interoperabili. Snowflake consente sia topologie basate su un singolo account che architetture multi-regione e multi-cloud e supporta l'integrazione di domini esterni o configurazioni per la collaborazione tra più aziende. La piattaforma di Snowflake sottostante, con Snowgrid a livello globale e il Marketplace Snowflake, è il tessuto connettivo che aiuta a evitare la creazione di nuovi silos di dati.

Inoltre, Snowflake offre un'ampia gamma di funzioni che aiutano le aziende a implementare i concetti di dati come prodotto e governance federata. In più, Snowflake si integra facilmente con un'ampia gamma di strumenti di terze parti che possono espandere le funzionalità della piattaforma. Il Data Cloud di Snowflake è un'eccellente scelta tecnologica per integrare i cambiamenti e i processi organizzativi necessari per trasformare efficacemente la data mesh. Per saperne di più su come Snowflake può aiutarti, visita [snowflake.com/data-mesh](https://snowflake.com/data-mesh)

# INFORMAZIONI SU SNOWFLAKE

Snowflake permette a ogni organizzazione di mobilitare i propri dati con il Data Cloud di Snowflake. 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 543 della classifica 2022 Forbes Global 2000 (G2K) al 31 ottobre 2022, utilizzano Snowflake Data Cloud per far crescere le loro aziende. Scopri di più sul nostro sito [snowflake.com](https://www.snowflake.com).



© 2022 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.