



PASSARE DALL'ETL ON-PREMISE ALL'ELT CLOUD-DRIVEN

Best practice per massimizzare il valore e l'efficienza delle pipeline di dati



SOMMARIO

- 2** Executive Summary
- 3** Terminologia e concetti di base
- 4** L'ascesa dei sistemi ELT
- 6** Definire una strategia di gestione dei dati versatile
- 7** Quando dovresti prendere in considerazione l'approccio ELT
- 7** Quale processo per la pipeline di dati
- 9** Elaborare i dati con Snowflake
- 10** Conclusione
- 11** Informazioni su Snowflake

EXECUTIVE SUMMARY

Le pipeline di dati progettate in passato erano concepite per i dati a bassa velocità, prevedibili e agevolmente categorizzati, provenienti da applicazioni aziendali on-premise. Si basano su processi di estrazione, trasformazione e caricamento (ETL, ovvero Extract, Transform, Load) per acquisire dati da varie fonti, trasformarli in un formato utile e caricarli in una destinazione target, come un data warehouse. Queste pipeline legacy funzionano bene con fonti di dati strutturati come le applicazioni aziendali, ma non sono più adeguate all'ampia varietà di tipi di dati e stili di ingestione che caratterizzano il panorama dei dati attuale.

Le moderne pipeline effettuano prima l'estrazione e il caricamento dei dati, per poi trasformarli quando saranno nella destinazione prevista: un ciclo noto come ELT (Extract, Load, Transform). Gli attuali sistemi ELT trasferiscono nel cloud i workload da trasformare, garantendo un livello di scalabilità ed elasticità nettamente superiore. Nei tradizionali ambienti on-premise, i processi ETL si contendono le risorse disponibili con altri workload in esecuzione sulla stessa infrastruttura. Con il sistema ELT, è possibile caricare i dati nella loro forma grezza e trasformarli in seguito in diversi modi, a seconda di come verranno utilizzati.

Con una pipeline ELT si possono caricare molti tipi di dati grezzi in un repository basato su cloud, come ad esempio una cloud data platform che migliora la velocità di ingestione, trasformazione e condivisione dei dati all'interno dell'organizzazione. Ciò consente

di eseguire i workload di trasformazione ad alta intensità di risorse direttamente nell'ambiente cloud, dove si potranno massimizzare la potenza di elaborazione e la capacità delle risorse scalabili.

Come vedremo nel seguito di questo ebook, il sistema ELT è una buona scelta nelle seguenti situazioni:

- **Se l'azienda deve gestire dati su larga scala**, l'ELT è in grado di elaborare rapidamente, nel cloud, grandi quantità di dati strutturati e non strutturati.
- **Per la sperimentazione analitica**, l'ELT massimizza le opzioni disponibili man mano che analisti e data scientist esplorano il potenziale dei dati, trasformandoli secondo necessità per progetti specifici.

- **Per pipeline di dati a bassa latenza**, l'ELT trasferisce i dati immediatamente, il che può risultare utile per le analisi a bassa latenza e i casi d'uso quasi in tempo reale.

Continua a leggere, per scoprire come massimizzare il valore delle tue pipeline di dati, utilizzando il metodo di trasformazione più adatto per ogni situazione e workload.



TERMINOLOGIA E CONCETTI DI BASE

L'ETL è un processo di integrazione software che prevede l'**estrazione** dei dati da varie fonti, la loro **trasformazione** in un server di staging e il successivo **caricamento** in una destinazione target, come ad esempio un data warehouse,

un data lake o una cloud data platform. Nei data warehouse tradizionali i dati sono mappati a un modello di dati relazionale. Possono essere sottoposti a processi di pulizia, arricchimento e trasformazione in un formato comune prima del caricamento nel database di destinazione.

La strutturazione e la trasformazione dei dati garantiscono un'analisi rapida ed efficiente con strumenti di business intelligence (BI) basati su SQL, ma limitano il modo in cui i dati possono essere utilizzati, dato che alcuni elementi dei dati grezzi si perdono nel processo di trasformazione. Nella maggior parte dei flussi di lavoro ETL, i dati

sono acquisiti dai database sorgente e trasferiti in un data warehouse. Un server di staging esegue la logica di trasformazione, che può comprendere filtraggio, mascheramento, arricchimento, mappatura, deduplicazione e integrazione di dati provenienti da più fonti.

I data engineer creano apposite pipeline di dati per orchestrare il movimento degli upload di dati in batch e per lo streaming continuo dei dati. Queste pipeline estraggono i dati da applicazioni, dispositivi e flussi di eventi. Le pipeline ETL trasformano i dati in un formato pronto per l'utilizzo aziendale come parte del flusso di lavoro ETL di base. Una soluzione praticabile quando le esigenze del business sono chiare. Tuttavia, per alcuni degli attuali workload più diffusi, come negli ambiti machine learning e data science, i requisiti di formato dei dati non sono sempre noti in anticipo. Ad esempio, i data scientist potrebbero decidere di conservare i dati allo stato grezzo (o meno elaborato), per convertirli poi in vari formati, adatti a diversi tipi di modelli, motori predittivi e scenari di analisi.



L'ASCESA DEI SISTEMI ELT

Le operazioni ETL di stampo tradizionale utilizzano un motore di elaborazione separato, spesso in esecuzione su server computazionali dedicati. Il database è modellato per custodire i dati in formati specifici e predefiniti, prima del caricamento degli stessi, in base alle esigenze di business downstream. Ad esempio, i dati si possono ordinare, riepilogare o parametrizzare per la visualizzazione rapida tramite dashboard, oppure raggruppare per i report finanziari mensili.

Queste procedure ETL possono funzionare bene in presenza di fonti di dati strutturati provenienti da applicazioni aziendali, come i sistemi di pianificazione delle risorse d'impresa (ERP), di gestione della supply chain (SCM) e di gestione delle relazioni con i clienti (CRM). Tuttavia, queste pipeline ETL legacy non ospitano agevolmente formati di dati più recenti, in volumi massivi, come i dati machine-generated dei sistemi Internet of Things (IoT), i dati in streaming provenienti dai social media, i dati di weblog dei siti internet e i dati di utilizzo mobile generati dalle app SaaS. Fanno un buon lavoro con l'ingestion di dati strutturati e in batch, ma si rivelano troppo rigide per la raccolta e l'ingestion di dati senza schema e semi-strutturati.

Le moderne pipeline sono concepite per grandi quantità di dati nei formati più recenti, per velocizzare le analisi effettuando prima l'estrazione

e il caricamento dei dati, per poi trasformarli una volta giunti a destinazione. Nella fase di trasformazione, i dati sono sottoposti a vari processi: standardizzazione, pulizia, mappatura e combinazione con dati provenienti da altre fonti. Le nuove pipeline di dati ELT sfruttano la potenza dei cloud data warehouse e delle cloud data platform, in grado di archiviare ed elaborare enormi quantità di dati a costi contenuti.

Oltre ai dati relazionali strutturati, le pipeline ELT possono eseguire l'ingestion di dati non strutturati, semi-strutturati e grezzi, per poi caricarli tutti quanti in una cloud data platform o un data lake. Lo staging dei dati non è necessario. I dati si possono conservare allo stato grezzo, per agevolare la sperimentazione e una rapida iterazione.

I VANTAGGI DELL'APPROCCIO ELT

Potenza:

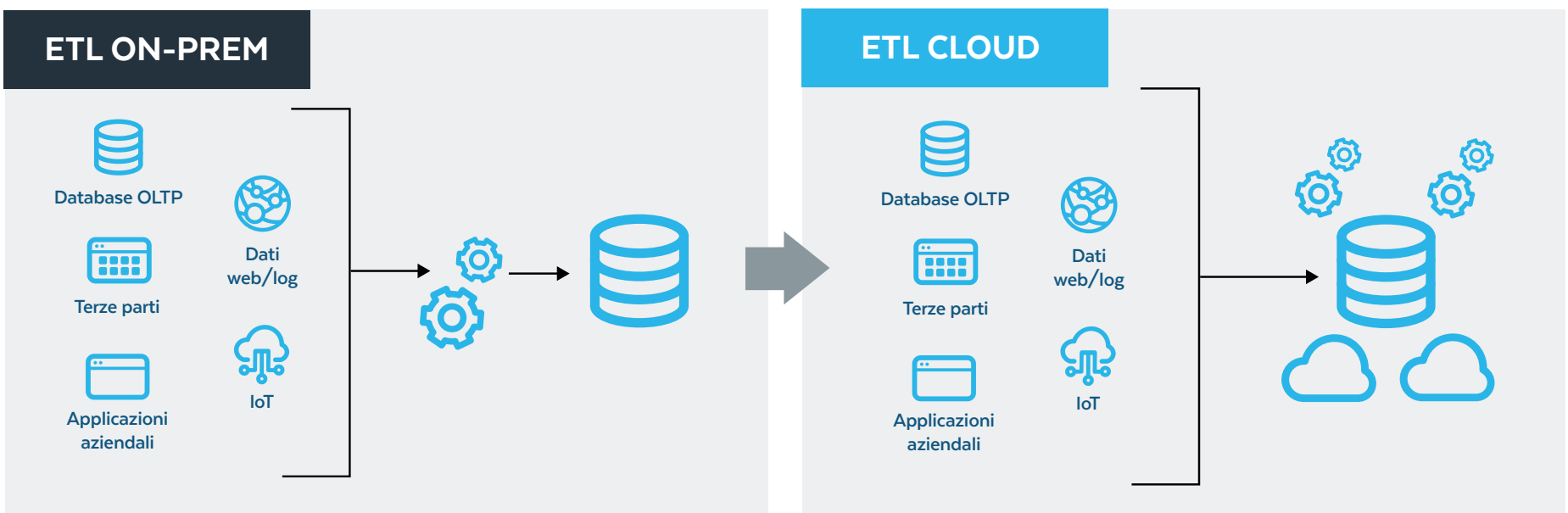
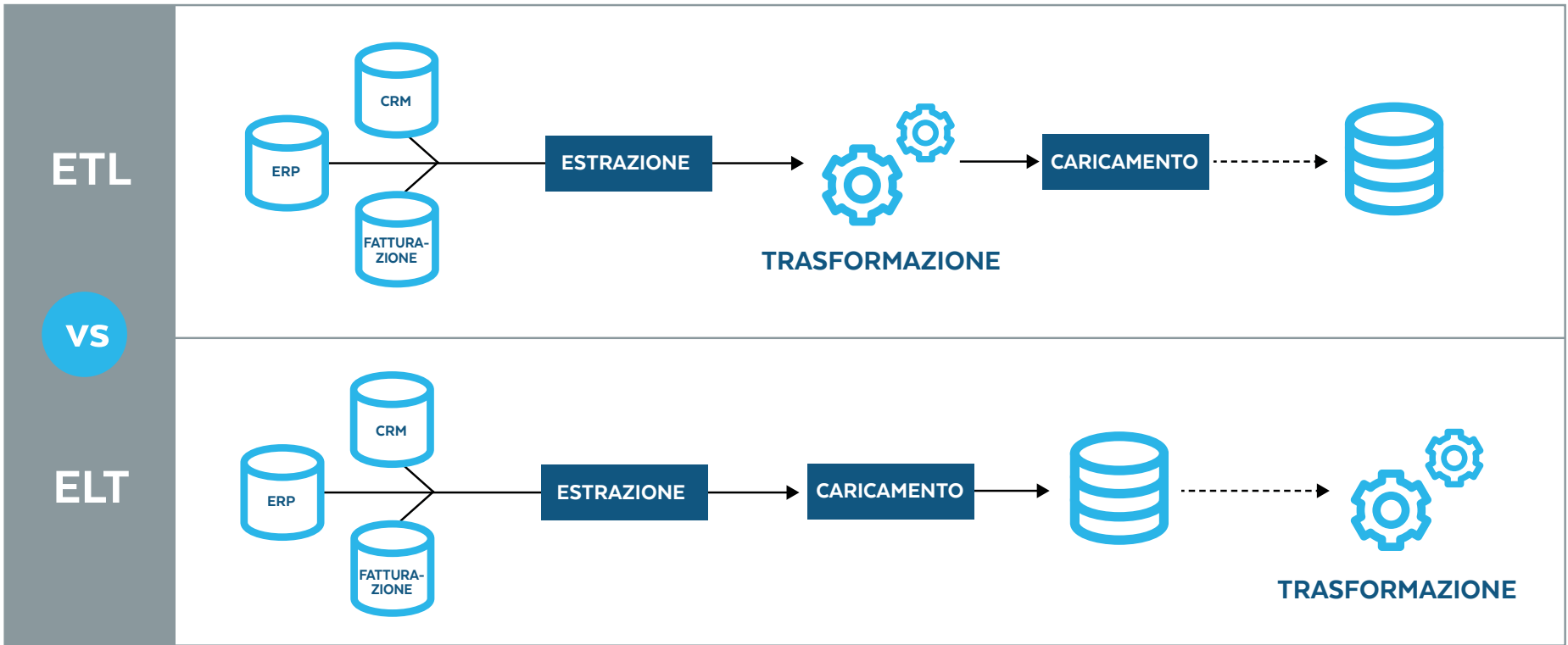
I repository basati su cloud offrono capacità di archiviazione quasi illimitate, supportate da server computazionali scalabili: tutto questo consente di restare al passo con crescenti volumi di dati.

Capacità:

Le pipeline ELT consentono l'ingestion di tutti i tipi di dati non appena i dati diventano disponibili, senza doverli trasformare in un formato specifico.

Flessibilità:

Trasforma solo i dati necessari per le analisi moment-to-moment, consentendo a più team di effettuare la trasformazione dei dati per report, dashboard, modelli di data science e altri task. Risultano così massimizzate le possibilità di utilizzo dei dati.



DEFINIRE UNA STRATEGIA DI GESTIONE DEI DATI VERSATILE

L'approccio ETL è un'opzione valida se i dati sono prevedibili, gestibili, e sono aggiornati regolarmente. Simili processi di batch ingestion sono comunemente utilizzati per gli application data che non necessitano di costanti aggiornamenti. Ad esempio, i dati dei punti vendita al dettaglio dovranno essere probabilmente aggiornati in un data warehouse alla fine di ogni giornata, per gestire i report quotidiani sugli incassi. Allo stesso modo, i dati dei clienti in un sistema CRM dovranno essere caricati sulla dashboard di un call center almeno una volta ogni ora, per riflettere le transazioni di vendita e assistenza attualizzate. O ancora, i dati sui consumi elettrici rilevati attraverso i contatori intelligenti hanno verosimilmente bisogno di essere aggiornati ogni 15 minuti, per supportare i programmi di fatturazione basati sul tempo di utilizzo.

Tuttavia, la situazione si può complicare rapidamente. Con i sistemi ETL e le architetture dati legacy, i dati provenienti da sistemi discreti sono abitualmente siloizzati in molti luoghi diversi. Ad esempio, ogni singola tipologia di dati può essere contenuta in un sistema unico, progettato e modellato per esigenze particolari. Ciò si traduce in numerosi repository diversi tra loro, che si possono trasformare alla svelta in un vero e proprio incubo a livello di manutenzione. On-premise o nel cloud, ogni applicazione di produzione crea il proprio silo di dati: dati di marketing in un sistema di marketing automation, dati di vendita in un sistema CRM, dati finanziari in un sistema ERP, dati di inventario in un sistema di gestione del magazzino, e così via. Ognuna di queste applicazioni può dipendere da strumenti ETL specializzati e procedure software uniche per la raccolta dei dati dai sistemi di produzione e la trasformazione degli stessi per le analisi.

Viste le moderne esigenze aziendali e di analisi, occorre consolidare tutti i dati in uno stesso posto, che sarà l'unica "fonte di verità", progettata per l'accesso universale da parte di vari gruppi di lavoro, applicazioni e strumenti.



QUANDO DOVRESTI PRENDERE IN CONSIDERAZIONE L'APPROCCIO ELT

La tecnologia ETL è tuttora diffusa nelle situazioni in cui i dati sono trasferiti da un sistema all'altro in modalità batch. Tuttavia, la maggior parte delle soluzioni ETL legacy non è in grado di gestire tutte le tipologie di dati. Funzionano bene per i dati strutturati delle applicazioni aziendali, ma non sono di certo l'ideale per i dati machine-generated prodotti dai sistemi IoT, i dati in streaming provenienti dai feed dei social media, i dati degli eventi in formato JSON e i dati di weblog su internet e delle app mobili.

Per decidere quale approccio utilizzare, ricorda queste linee guida:

- I processi ETL funzionano bene con i dati relazionali che devono conservare una struttura tabulare.
- I processi ELT sono il metodo migliore per i dati semi-strutturati da mantenere nel formato nativo o grezzo fino alla definizione dei casi d'uso di analisi specifici.

Altre considerazioni riguardano l'effettiva quantità di dati da elaborare e la rapidità con cui devono essere preparati per l'analisi downstream. I processi di trasformazione richiedono molti cicli di calcolo. Con l'ELT, la scalabilità automatica fornisce istantaneamente le risorse necessarie per supportare ogni operazione. L'utilizzo di un processo ELT consente di sfruttare le risorse illimitate del cloud per elaborare e trasformare i dati in modo rapido ed efficiente. Inoltre, riduce al minimo lo spostamento dei dati, che si possono elaborare nella posizione in cui risiedono, anziché trasferirli su un server o un meccanismo di archiviazione indipendente.

Occorre determinare dove si eseguirà il motore di elaborazione, quali risorse sono disponibili a livello di infrastruttura e di quali prestazioni si ha bisogno. Hai problemi di scalabilità o utenti concorrenti, come una capacità limitata del server? Indipendentemente dal fatto che i dati siano prodotti da sistemi di elaborazione delle transazioni online (OLTP), interazioni con siti web, app SaaS, sensori di apparecchiature o flussi provenienti dai social media, i data engineer devono sviluppare apposite pipeline per acquisire tali dati, inserirli in un repository e renderli accessibili agli utenti aziendali. In molti casi, le operazioni sulla pipeline di dati vengono migliorate sfruttando la potenza di elaborazione dei database di destinazione situati nel cloud.

QUALE PROCESSO PER LA PIPELINE DI DATI

Scegli l'approccio ETL quando:

- Il volume totale dei dati da elaborare è relativamente limitato
- I database di origine e di destinazione richiedono tipologie di dati diverse
- Elaborate principalmente dati strutturati

Scegli l'approccio ELT quando:

- Devi elaborare grandi volumi di dati
- I database di origine e di destinazione sono dello stesso tipo
- I dati sono semi-strutturati o non strutturati



CASE STUDY DI UN CLIENTE SNOWFLAKE

L'organizzazione: Paciolan è un'azienda leader nello sviluppo di soluzioni tecnologiche per ticketing, fundraising, marketing e analytics con oltre 500 organizzazioni clienti nel settore del live-entertainment che vendono complessivamente oltre 120 milioni di biglietti all'anno.

Il problema: per convertire i dati semi-strutturati in dati relazionali, Paciolan scrive codice ETL proprietario per l'analisi e la normalizzazione dei dati. Tuttavia, cinquantamila record possono trasformarsi in 1 milione di righe in un data warehouse on-premise. Il processo ETL, con quasi 100 GB di dati al giorno, richiedeva da 30 a 60 minuti per essere

completato. Le risorse limitate impedivano agli analisti di riassumere e consolidare efficacemente i dati.

La soluzione: Paciolan ora archivia i dati JSON semi-strutturati come tipo di dati VARIANT nella piattaforma Snowflake. Utilizza Snowflake sia come data lake, sia come data warehouse tramite Data Vault, un approccio architetturale che prevede uno specifico schema di progettazione del modello dati per supportare un data warehouse aziendale moderno e agile.

I risultati: il processo ETL, che richiedeva fino a un'ora nel data warehouse legacy, con la pipeline di dati Snowflake adesso viene completato in pochi minuti. Gli sviluppatori possono utilizzare semplici script Python per inserire le istruzioni in modo dinamico.

I vantaggi

- La separazione tra archiviazione e capacità di calcolo garantisce stabilità delle prestazioni e visibilità dei costi
- L'elasticità immediata consente una potenza di calcolo quasi illimitata per supportare, virtualmente, qualsiasi numero di utenti
- Il supporto per l'archiviazione dei dati semi-strutturati come tipo di dati variant consente di ottenere insight sui dati ancor più approfonditi

“Abbiamo confrontato i numeri prima e dopo. Abbiamo scoperto che con Snowflake si otteneva una riduzione del 90% del codice utilizzato per il processo ETL. È una grande vittoria per noi.”

Ashkan Khoshcheshmi
Principal Software Engineer
Paciolan



ELABORARE I DATI CON SNOWFLAKE

La piattaforma Snowflake comprende funzionalità per la pipeline di dati flessibili e scalabili, come parte del servizio base. In Snowflake puoi inserire direttamente i dati grezzi, senza necessità di creare una pipeline per trasformare i dati in un formato diverso. Snowflake esegue queste trasformazioni automaticamente, riducendo al minimo i costi di storage ed elaborazione.

Inoltre, Snowflake semplifica la gestione dei dati, eliminando i silos: non dovrai pertanto conservare più copie dei dati per le diverse applicazioni downstream. Mantiene la forma originale dei dati grezzi, applicando al tempo stesso, in maniera trasparente, tecniche di archiviazione altamente ottimizzate, in modo che le analisi e le trasformazioni dei dati possano funzionare eccezionalmente bene.

Soprattutto, Snowflake è progettato per sfruttare appieno le caratteristiche uniche del cloud. Si basa su un'architettura dati condivisa multi-cluster che separa la capacità di calcolo dallo storage, consentendo di trasformare i dati su scala. Ciascun tipo di risorsa è scalabile in modo indipendente, per soddisfare le specifiche esigenze di ogni applicazione.

La piattaforma Snowflake è stata sviluppata attorno a un robusto motore di elaborazione, progettato per gestire tutti i tipi di workload, senza degrado delle prestazioni. Consente ad esempio l'ingestion dei dati tramite una pipeline di data engineering e il contemporaneo training di un modello di machine learning per l'utilizzo degli stessi dati. Il servizio di

pipeline scalabile è in grado di eseguire l'ingestion dei dati in modo continuativo, senza influire sulle prestazioni di questi o altri workload. I data engineer possono decidere quanta potenza di calcolo allocare a ciascun processo di ingestion o consentire al sistema di scalare automaticamente.

Snowflake consente ai data engineer di sviluppare pipeline di dati con un'ampia scelta di linguaggi e strumenti di integrazione per la gestione del flusso di dati e supporta una vasta gamma di stili di ingestion, come l'integrazione in batch o in streaming con Apache Kafka. Inoltre, con Snowflake è possibile caricare in modo semplice ed efficiente molti tipi di dati utilizzando l'SQL standard, la lingua franca nel campo dei database.



CONCLUSIONE

La crescita costante del volume, della varietà e della velocità dei dati richiede nuovi tipi di pipeline e motori di elaborazione basati su cloud più evoluti per acquisire i dati e utilizzarli in modo efficiente.

I processi ETL sono spesso gestiti da server on-premise con capacità fissa, larghezza di banda limitata e un numero finito di cicli CPU. I moderni workload di integrazione dei dati migliorano sfruttando la potenza di elaborazione dei database cloud e delle cloud data platform, scalabili secondo necessità.

Per trarre vantaggio dalle risorse cloud, un numero crescente di organizzazioni sta sviluppando pipeline di dati in grado di estrarre e caricare i dati in un database cloud, per poi trasformarli una volta che saranno arrivati a destinazione. Questo approccio, noto come ELT, è più rapido rispetto ai tradizionali processi ETL, poiché sfrutta la potenza dei moderni motori di elaborazione limitando gli spostamenti di dati allo stretto indispensabile.

I processi ELT trasferiscono nel cloud i workload di trasformazione ad alta intensità di risorse per due motivi fondamentali:

1. **Sfruttare le risorse quasi illimitate del cloud per elaborare e trasformare i dati in modo rapido ed efficiente**
2. **Mantenere i dati allo stato grezzo fino a quando non saranno chiare le effettive esigenze del business**

Quando è possibile, preferisci sempre l'approccio ELT all'ETL per trasferire i processi di trasformazione ad alta intensità di risorse su una piattaforma di destinazione cloud-based. L'approccio ELT semplifica il funzionamento delle pipeline, minimizza gli spostamenti dei dati, riduce il numero dei silos e massimizza le possibilità di utilizzo finale dei dati.

Per saperne di più sulle soluzioni Snowflake per le pipeline di dati, visita snowflake.com/it/data-cloud/workloads/data-engineering/.





INFORMAZIONI SU SNOWFLAKE

Snowflake permette a ogni organizzazione di mobilitare i propri dati grazie al Data Cloud. I clienti utilizzano il Data Cloud per unificare i dati contenuti nei silos, esplorare e condividere i dati in totale sicurezza, potenziare le applicazioni basate sui dati, ed eseguire diversi workload di AI/ML e 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 590 della classifica 2022 Forbes Global 2000 (G2K) al 30 aprile 2023, utilizzano il Data Cloud di Snowflake per far crescere le loro aziende.

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



©2021 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.