Democratizzare l’accesso ai dati tramite una Data Platform self-service utilizzando AWS LakeFormation – Parte 2

12 Marzo 2025 - 13 min. read

Data Ingestion | Data Platform | Medallion architecture

In questa serie di articoli, stiamo descrivendo come creare e strutturare correttamente una Data Platform self-service per la democratizzazione dei dati analitici su AWS. Dall'acquisizione e archiviazione dei dati, attraverso strumenti di elaborazione per creare dati preziosi per analisi, visualizzazioni e reportistica.

Ci concentriamo anche sulla governance dei dati, sulla loro identificazione e sulla collaborazione, con un'attenzione particolare alla sicurezza e al controllo degli accessi.

Segui questo articolo per imparare come democratizzare l'accesso ai dati attraverso la tua Data Platform self-service. Vedremo come garantire la governance, strutturando correttamente dati, accessi e visibilità, utilizzando AWS LakeFormation. Non dimenticare di tenere d'occhio il sito web in attesa della parte 3!

Questo articolo è un seguito alla descrizione delle Data Platform e delle relative pipeline di dati, basandosi e costruendo su questi concetti. Se stai ancora cercando di familiarizzare con questi concetti, o hai bisogno di un ripasso, ecco la Parte 1.

TL;DR

Inserisci le tue fonti di dati nei bucket S3 e registra le posizioni dei dati all'interno di AWS LakeFormation. Cataloga i dati con database, tabelle e colonne. Definisci e associa LF-Tags a queste risorse del catalogo per eseguire un controllo degli accessi

basato sugli attributi (ABAC). Definisci ruoli e concedi loro autorizzazioni basate sui tag per abilitare l'accesso ai dati. Crea un amministratore con permessi assegnabili su aree specifiche e utilizza i tag per la identificazione dei dati al fine di democratizzare e ottenere un accesso self-service ai dati.

La Sfida della Democratizzazione dei Dati

Nel mondo odierno guidato dai dati, le organizzazioni affrontano un paradosso critico: nuotano in vasti oceani di dati, sebbene la maggior parte di esse fatichi a utilizzare efficacemente questa preziosa risorsa.

Gli approcci tradizionali alla gestione dei dati tendevano ad organizzarli in strutture separate e sconnesse come i silos. In questi approcci, ogni silo è solitamente accessibile solo dal proprio dipartimento tecnico, creando diversi problemi lungo il percorso.

Le sfide della democratizzazione dei dati vanno oltre le limitazioni tecniche. Questa separazione in silos crea barriere complesse che impediscono agli analisti un accesso diffuso ai dati, come il dover presentare richieste dispendiose in termini di tempo ai team IT o ai team dei dati anche per l'accesso ai dati più basilari. Gli utenti operano con informazioni incomplete, avendo grandi difficoltà a vedere il "quadro generale" e il potenziale vantaggio competitivo del processo decisionale basato sui dati rimane irrealizzato.

Molte aziende si trovano intrappolate in un ciclo di gestione manuale degli accessi, dove le richieste di accesso ai dati richiedono molteplici approvazioni, configurazioni complesse dei permessi e manutenzione continua. Questo non solo crea un significativo onere amministrativo, ma **rallenta anche il potenziale di innovazione**.

L'architettura del data lake aiuta a risolvere questa sfida concentrando tutti i dati in un unico luogo. Chiunque necessiti di accedere ai dati sa dove cercare. Ma non è tutto oro quel che luccica! Aggregare tutti i dati in un unico posto crea una sfida nuova, ma diversa: la gestione degli accessi utente. Anche se ora potenzialmente tutti possono avere accesso ai dati, è sicuro?

Le organizzazioni devono bilanciare simultaneamente due priorità contrastanti: abilitare un ampio accesso ai dati mantenendo al contempo rigorosi protocolli di governance e sicurezza. Il rischio di esporre informazioni sensibili, unito ai requisiti di

conformità come GDPR, CCPA e normative specifiche del settore, crea un significativo sovraccarico nella gestione delle autorizzazioni sui dati.

È qui che AWS LakeFormation può diventare uno strumento molto utile!

Cos'è AWS LakeFormation?

AWS Lake Formation è un servizio completamente gestito che semplifica la creazione, la sicurezza e la gestione dei data lake.

Nella sua essenza, il servizio semplifica il processo tradizionalmente complesso e dispendioso in termini di tempo di consolidamento dei dati provenienti da molteplici fonti in un repository unificato e sicuro - il data lake - in pochi giorni invece che in mesi/anni. A differenza degli approcci tradizionali di gestione dei dati che richiedono un'ampia configurazione manuale e una complessa configurazione dell'infrastruttura, AWS LakeFormation automatizza compiti critici come l'acquisizione dei dati, la catalogazione dei metadati e il controllo degli accessi. È una piattaforma centralizzata che astrae dalle complessità tecniche, permettendo a ingegneri di dati, analisti e leader aziendali di concentrarsi su ciò che conta davvero: estrarre valore reale dai dati.

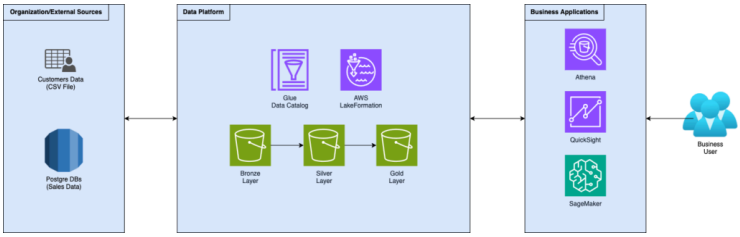
Inoltre, AWS LakeFormation fornisce solide capacità di governance e sicurezza, caratteristiche essenziali per la governance dei dati nelle imprese data-driven. Il servizio offre controlli di accesso granulari basati sugli attributi che consentono alle organizzazioni di definire politiche di accesso ai dati precise a livello di database, tabella, colonna e persino riga. Questo significa che le aziende possono implementare meccanismi di sicurezza dettagliati che garantiscono la protezione delle informazioni sensibili pur essendo in grado di ottenere la democratizzazione dei dati. Grazie all'integrazione diretta con altri servizi AWS come Amazon S3, AWS Glue e Amazon Athena, AWS LakeFormation crea un ecosistema completo che supporta l'intero ciclo di vita dei dati, dall'acquisizione e trasformazione dei dati grezzi all'analisi e alla visualizzazione. La sua capacità di centralizzare la gestione dei metadati, automatizzare la scoperta dei dati e fornire una sicurezza coerente su diverse fonti di dati lo rende uno strumento fondamentale per le imprese che cercano di sfruttare le proprie risorse di dati in modo efficiente e sicuro.

Governare il Data Lake

Ora che abbiamo descritto le sfide e gli strumenti, sporchiamoci le mani e mettiamoli in azione!

Se hai letto il primo di questa serie di articoli, sai già con cosa stiamo lavorando ma, per mettere tutti sulla stessa pagina, ecco una brevissima panoramica della configurazione.

Agendo come data engineer in un'azienda fittizia che aiuta i suoi clienti ad aumentare i loro ricavi, hai creato una data platform, seguendo l'architettura standard "medallion". Hai sviluppato logiche di acquisizione e trasformazione per raccogliere i dati e spostarli attraverso i livelli sempre più raffinati della data platform.



L'azienda ora ti chiede di governare la data platform, rendendo i dati accessibili ai team interni e ai clienti.

I clienti vogliono solo vedere e interrogare i loro dati, mentre i team interni hanno bisogno di visualizzare i dati e utilizzarli per addestrare modelli di Machine Learning che li aiutino a supportare i clienti nel raggiungimento dei loro obiettivi.

Inoltre, devi tenere d'occhio l'accesso ai dati e la sicurezza: i clienti devono vedere solo i loro dati!

In più, i dati dei clienti contengono informazioni personali (PII) che non sono utili per i team interni e non dovrebbero essere visibili a loro.

Ingestion dei Dati

Abbiamo già tutti i dati grezzi inseriti all'interno del bucket del livello bronzo, tuttavia, ecco un rapido suggerimento che potrebbe essere utile per alcuni dei lettori che stanno cercando di implementare l'ingestion.

AWS LakeFormation offre blueprint per importare dati da database relazionali, CloudTrail e log dei load balancer. I blueprint sono template CloudFormation predefiniti che creano tutte le risorse necessarie per eseguire l'ingestion delle tue fonti dati. Di fatto creano un workflow Glue, composto da Glue job e crawler che inseriscono i dati all'interno dei tuoi bucket S3 e aggiornano il Glue Data Catalog.

Registrare le Location del Data Lake

Prima di tutto, dobbiamo far conoscere ad AWS LakeFormation gli asset che compongono il nostro data lake. Per farlo, dobbiamo registrare le location S3. Possiamo registrare bucket o percorsi specifici al loro interno. Seguendo l'architettura medallion, abbiamo creato 3 bucket:

- Livello Bronzo: lakeformation-app-landing
- Livello Argento: lakeformation-app-processed
- Livello Oro: lakeformation-app-presentation

Un suggerimento qui è quello di strutturare i dati all'interno dei bucket tenendo a mente chi ha bisogno di accedere a quali dati e che tipo di query devono essere eseguite su quei dati. Questo aiuterà a strutturare facilmente i permessi per accedere ai dati e rendere le query veloci ed efficienti.

Nel nostro esempio, abbiamo strutturato i dati dividendoli tra le unità interne che hanno bisogno di accedervi. Nel nostro caso: marketing e vendite.

Eseguiamo questa operazione per tutti i bucket e i percorsi che abbiamo definito.

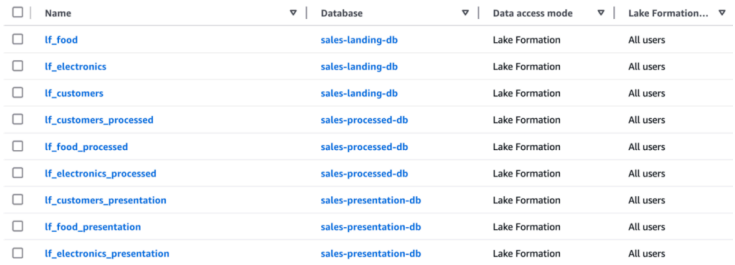
Catalogazione dei Dati

Ora che abbiamo definito i componenti del nostro data lake, è il momento di catalogare i dati. Iniziamo creando i database sopra le posizioni del data lake. Il database ci permette di creare strutture logiche, tabelle e viste, sul data lake in modo che gli utenti possano facilmente accedere e interrogare i dati tramite SQL.

Dato che stiamo facendo un esempio di test con pochissime tabelle, abbiamo creato solo un database per ogni livello dell'architettura medallion. L'idea qui è di mostrare la gestione dell'accesso ai dati utilizzando AWS LakeFormation, quindi, abbiamo diviso i dati in due livelli:

- Livello di contenuto sensibile: abbiamo una tabella con PII (informazioni personali identificabili dei clienti) e tabelle con dati non sensibili (vendite)
- Separazione dei team interni: abbiamo creato 2 tabelle di vendite (elettronica e alimentari), imitando due diversi team interni, uno associato al cliente che vende alimentari e l'altro con il negozio di elettronica.

Nel primo capitolo di questa serie di articoli, abbiamo costruito la pipeline di dati che esegue l'ingestion e il raffinamento attraverso i vari livelli della data platform, quindi, abbiamo già i dati al loro posto. Utilizziamo i Glue Crawler per identificare e catalogare automaticamente i dati all'interno del Glue Data Catalog, che è integrato con AWS LakeFormation. Dopo aver eseguito i crawler, ecco le tabelle su cui lavoreremo:



Name	Database	Data access mode	Lake Formation...
lf_food	sales-landing-db	Lake Formation	All users
lf_electronics	sales-landing-db	Lake Formation	All users
lf_customers	sales-landing-db	Lake Formation	All users
lf_customers_processed	sales-processed-db	Lake Formation	All users
lf_food_processed	sales-processed-db	Lake Formation	All users
lf_electronics_processed	sales-processed-db	Lake Formation	All users
lf_customers_presentation	sales-presentation-db	Lake Formation	All users
lf_food_presentation	sales-presentation-db	Lake Formation	All users
lf_electronics_presentation	sales-presentation-db	Lake Formation	All users

Un altro consiglio è quello di dedicare del tempo all'aggiunta di metadati e descrizioni a database, tabelle e colonne per migliorare la scopribilità. Ad esempio, puoi aggiungere descrizioni aziendali in modo che gli utenti business possano utilizzare la propria terminologia per trovare i dati giusti, riducendo drasticamente il tempo speso nella scoperta dei dati.

Un'altra best practice è associare i tag ai dati all'interno di AWS LakeFormation. Gli **LF-Tags** (Lake Formation Tags) sono coppie chiave-valore associabili a database, tabelle e colonne per descrivere caratteristiche come la sensibilità dei dati, il dominio aziendale o qualsiasi altra entità rilevante.

Utilizzeremo questi tag successivamente per implementare il controllo degli accessi basato sugli attributi (ABAC), un approccio molto più scalabile rispetto al tradizionale controllo degli accessi basato sui ruoli.

Governance dei Dati

Ora che abbiamo tutte le fonti nel nostro data lake, iniziamo a costruire le basi per eseguire l'amministrazione del data lake e concedere l'accesso ai dati in modo semplice.

Come menzionato in precedenza, iniziamo definendo la nostra strategia di tag con LF-Tags per implementare il controllo degli accessi basato sugli attributi (ABAC) e gestire l'accesso ai dati su larga scala. Questo passaggio è cruciale per la democratizzazione dei dati poiché utilizzeremo questi tag sia per concedere l'accesso a tali dati sia per consentire agli utenti di auto-scoprire i dati a loro necessari.

Nel nostro esempio, abbiamo sviluppato questa semplice strategia:

- Area: Marketing, Sales
- Domain: Customers, Food, Electronics
- Sensitivity: Public, Private, PII

Dopo aver creato gli LF-Tags con i loro set di valori consentiti, possiamo iniziare ad assegnarli a database, tabelle o viste, e persino a colonne specifiche. Notare che gli LF-Tags vengono propagati nelle strutture di livello inferiore, quindi tutti i tag associati a un database saranno associati a tutte le sue tabelle. Allo stesso modo, tutti i tag associati a una tabella/vista saranno associati a tutte le sue colonne.

Il vero potere degli LF-Tags, rispetto agli approcci tradizionali, appare quando si implementano modelli di accesso complessi, essendo in grado di seguire il principio least-privilege con straordinaria precisione. I data steward possono creare politiche di accesso basate su tag che concedono automaticamente permessi agli utenti che possiedono tag corrispondenti nelle loro espressioni LF-Tag. Alcuni esempi rapidi basati sui nostri dati di esempio:

- Una policy che stabilisce che gli utenti con tag *UserRole=FoodAnalyst* possono accedere a qualsiasi dato etichettato con *Domain=Food* e *Sensitivity=Public*.
- Una policy che stabilisce che gli utenti con tag *UserRole=MarketingAnalyst* possono accedere a qualsiasi dato etichettato con *Sensitivity=PII*. Questo framework di classificazione è molto utile per identificare rapidamente tutte le risorse contenenti PII, o altri tipi di dati regolamentati, in tutto il data lake. Possiamo usarlo per nascondere alcune colonne ad utenti che non hanno il permesso di vedere questo tipo di informazioni. Inoltre, permette una reportistica completa per audit e conformità.

Questo approccio riduce drasticamente il sovraccarico operativo man mano che il data lake cresce. Quando vengono aggiunti nuovi dataset, i loro amministratori devono solo applicare i tag appropriati e i corretti controlli di accesso saranno applicati automaticamente. Allo stesso modo, quando un nuovo utente si unisce, gli amministratori del data lake devono solo assegnare le appropriate espressioni LF-Tag e l'accesso a tutti i dati rilevanti all'interno dell'organizzazione sarà concesso.

Ruoli e Amministrazione dei Dati

I dati sono ora categorizzati con tag e abbiamo sviluppato politiche di tag per definire come accedervi. È il momento di definire le entità che accederanno ai dati e iniziare a concedere i permessi.

Innanzitutto, definiamo le entità che accederanno ai nostri dati. L'idea qui è quella di creare ruoli che riflettano le funzioni aziendali all'interno dell'organizzazione. Questo approccio è molto efficace poiché quando una nuova persona si unisce al progetto, deve semplicemente essere assegnata al ruolo appropriato piuttosto che richiedere configurazioni di permessi personalizzate, riducendo significativamente il carico amministrativo.

Nel nostro caso, abbiamo definito:

- Due ruoli amministrativi: uno per l'area marketing e uno per l'area vendite
- Un ruolo di analista: per il dominio alimentare dell'area vendite

L'idea centrale qui è quella di delegare l'amministrazione del database ai ruoli amministrativi locali in modo che possano concedere i permessi autonomamente, promuovendo un approccio democratizzato ai dati.

Finalmente, è il momento di concedere i permessi!

Accesso ai Dati Self-Service

Ora che abbiamo posizionato tutti i pezzi necessari, abbiamo le basi per ottenere un accesso ai dati in modalità self-service.

Avere tutti i dati strutturati e organizzati in AWS LakeFormation, con il contenuto adeguatamente descritto tramite tag, in modo molto preciso, fino al livello di colonna, permette agli utenti di cercare facilmente i dati necessari e iniziare a estrarne valore.

Come amministratori di AWS LakeFormation, iniziamo delegando l'amministrazione dei dati di ogni area dell'organizzazione agli amministratori locali. Come menzionato in precedenza, definiamo due amministratori, uno per l'area marketing e uno per le vendite, utilizzando le espressioni LF-Tag:

- *Area = Marketing*
- *Area = Sales*

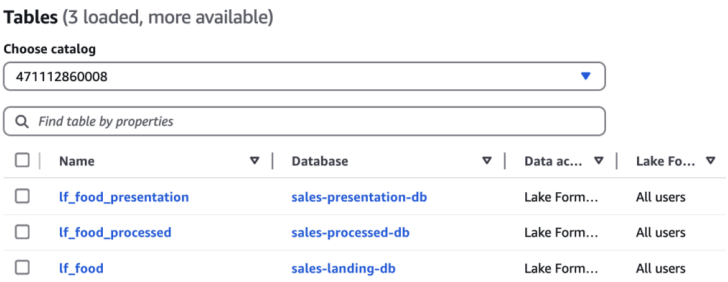
Con queste semplici espressioni, selezioniamo tutti i database, le tabelle e le colonne che sono taggati con quei valori di LF-Tag.

Poiché stiamo creando ruoli amministratore, concediamo autorizzazioni “super” sia sui database che sulle tabelle per una completa delega. Inoltre, molto importante per lo stesso motivo, consentiamo anche la concessione dei permessi di lettura a loro volta verso altre utenze. In questo modo potranno fornire autonomamente gli accessi al proprio team.

Qui vediamo la vera potenza del modello di autorizzazioni di AWS LakeFormation che, al contrario delle policy IAM, fornisce controlli specifici per i dati che operano indipendentemente dallo strato di archiviazione sottostante. Questo è un vantaggio cruciale poiché le policy di accesso ai dati rimangono coerenti indipendentemente da come gli utenti accedono ai dati. Che si tratti di una query su una tabella tramite Athena, un'analisi con SageMaker o una dashboard utilizzando QuickSight, le autorizzazioni di AWS LakeFormation rimangono le stesse e si applicano in modo coerente.

Gli amministratori locali di area possono ora vedere tutte le tabelle associate alle rispettive aree all'interno dell'organizzazione. Per completare l'esempio, possiamo utilizzare l'amministratore dell'area vendite per concedere l'accesso ai dati di vendita dei prodotti alimentari al suo team, rappresentato dal ruolo di analista dati. In questo modo, l'analista sarà in grado di vedere e utilizzare i dati di vendita dei prodotti alimentari per eseguire analisi e creare dashboard per la visualizzazione.

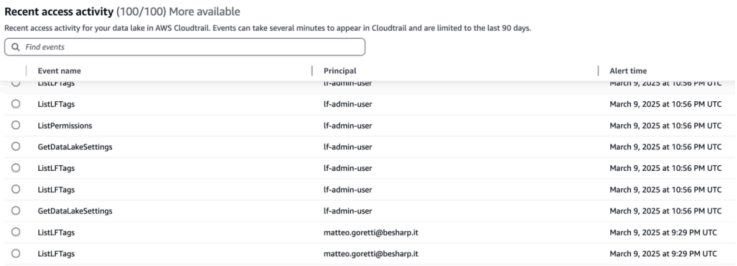
Abbiamo finalmente democratizzato l'accesso ai dati. Gli utenti possono ora liberamente scoprire i dati utilizzando i tag e accedervi!



Monitoraggio e Audit

Prima di concludere, diamo un'occhiata alle funzionalità di monitoraggio e audit di AWS LakeFormation.

AWS LakeFormation offre una dashboard centralizzata che i team di sicurezza possono utilizzare per esaminare tutti gli eventi di accesso ai dati. Questa centralizzazione consente una governance più efficace, semplifica la conformità e permette risposte più rapide agli audit. Ecco un esempio di immagine della dashboard:



A prima vista, la dashboard riporta l'evento, insieme all'utente e all'ora dell'evento. Ogni evento ha una descrizione dettagliata, che è il suo log all'interno di CloudTrail.

Punti principali e piccoli spoiler

In questo articolo, abbiamo esplorato come AWS LakeFormation possa trasformare l'accesso ai dati nelle organizzazioni attraverso una piattaforma di dati self-service.

Dalla configurazione, registrazione delle location dei dati, catalogazione completa con metadati aziendali e LF-Tags per il controllo degli accessi basato sugli attributi, le organizzazioni possono ottenere una vera democratizzazione dei dati mantenendo al contempo una solida sicurezza.

La potenza di AWS LakeFormation risiede nella sua capacità di definire permessi granulari a livello di database, tabella, colonna e riga, permettendo agli amministratori di delegare il controllo degli accessi ai data steward specifici per area. Questo approccio riduce significativamente il sovraccarico amministrativo assicurando al contempo che gli utenti possano scoprire e accedere solo ai dati di cui hanno bisogno.

Attraverso una corretta implementazione di strutture basate sui ruoli allineate con le funzioni aziendali, AWS LakeFormation crea una base per un processo decisionale basato sui dati in tutta l'azienda.

Ora che sappiamo come democratizzare i dati con AWS LakeFormation, nel prossimo capitolo di questa serie di articoli esploreremo i servizi che si basano su AWS LakeFormation, come DataZone.

Hai mai provato a democratizzare l'accesso ai dati per conto tuo?

Faccelo sapere nei commenti!

About Proud2beCloud

Proud2beCloud è il blog di beSharp, APN Premier Consulting Partner italiano esperto nella progettazione, implementazione e gestione di infrastrutture Cloud complesse e servizi AWS avanzati. Prima di essere scrittori, siamo Solutions Architect che, dal 2007, lavorano quotidianamente con i servizi AWS. Siamo innovatori alla costante ricerca della soluzione più all'avanguardia per noi e per i nostri clienti. Su Proud2beCloud condividiamo regolarmente i nostri migliori spunti con chi come noi, per lavoro o per passione, lavora con il Cloud di AWS. Partecipa alla discussione!

Matteo Goretti

DevOps Engineer @ beSharp. Appassionato di Cloud Computing e Intelligenza Artificiale, in particolare, Machine Learning e Deep Learning. Amo il trekking e la natura in generale. Mi rilasso con la mia chitarra, giocando ai videogames o guardando serie TV.
