Costruire un Data Lake su AWS con AWS Lake Formation

16 Marzo 2021 - 11 min. read

Amazon S3 | AWS Identity and Access Management (IAM) | AWS Lake Formation | Data and Analytics | Data Lake | Data Security and Governance | MLOps

Introduzione

Sfruttare i dati disponibili (Big Data) è diventato un obiettivo di primaria importanza per tantissime aziende negli ultimi decenni. In questi anni infatti, l'avvento del Cloud Computing ha democratizzato l'accesso a risorse IT estremamente potenti, risultato che ha portato all'eliminazione dei costi e delle difficoltà di gestire l'infrastruttura necessaria al funzionamento di un data center on-premise.

Il Cloud Computing aiuta le aziende a utilizzare i dati in modo efficace, riducendo i costi di ingegneria grazie alla potenza e versatilità dei suoi servizi managed.

Promuovendo l'utilizzo di infrastrutture on-demand, rende più facile ripensare, re-ingegnerizzare, e re-architettare un data lake per poter esplorare nuovi casi.

Essendo dunque i dati l'elemento determinante per le decisioni di business, gestirli in modo efficace diventa una priorità.

Tra i tanti modi per farlo, il concetto di data lake, ovvero un repository, scalabile, low-cost, per centralizzare i dati provenienti da fonti diversification, è diventato uno dei più apprezzati. Esso permette agli utenti di salvare i dati as-is senza un processo di strutturazione a priori, e può essere utilizzato per effettuare analisi di vario tipo, ottenere insights sui dati, e guidare con più efficacia le decisioni di business.

Ma creare un data lake da zero non è un compito facile: è necessario eseguire diverse operazioni manuali, che rendono il processo complesso e, cosa più importante, estremamente dispendioso dal punto di vista del tempo impiegato. I dati, di solito, provengono da fonti disparate e per questo, vanno monitorati con cautela.

Inoltre, gestire una tale quantità di dati richiede l'applicazione di diverse procedure per evitare leaks e buchi di sicurezza, il che significa mettere in atto un sistema di gestione delle policy di accesso, abilitare la cifratura dei dati sensibili e, naturalmente, gestirne le chiavi.

Senza applicare le giuste scelte riguardo la tecnologia, l'architettura, la qualità dei dati e la governance degli stessi, un data lake può diventare velocemente un sistema caotico ed isolato, difficile da utilizzare, da mantenere, e spesso isolato.

Per fortuna, il Cloud di AWS viene in nostro aiuto grazie ai molti servizi disegnati appositamente per gestire i data lake, in particolare AWS Glue e S3.

Per questo articolo, assumiamo che il lettore sia già familiare con i concetti di servizi AWS e che conosca le peculiarità dietro a AWS Glue e S3. Qualora non fosse così, invitiamo a leggere le nostre ultime "stories" su ingesting data for Machine Learning workloads e managing complex Machine Learning projects via Step Functions.

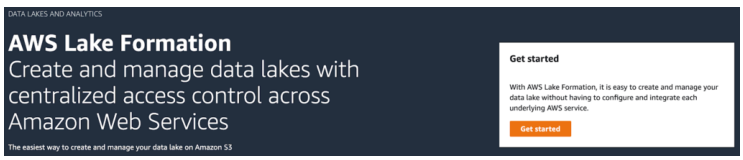
Andremo ad esplorare come costruire un semplice data lake con Lake Formation. Quindi, ci concentreremo sugli aspetti di sicurezza e governance, ed esploreremo i vantaggi che questo servizio offre, rispetto al semplice utilizzo di AWS Glue.

Partiamo!

Setup base di Lake Formation

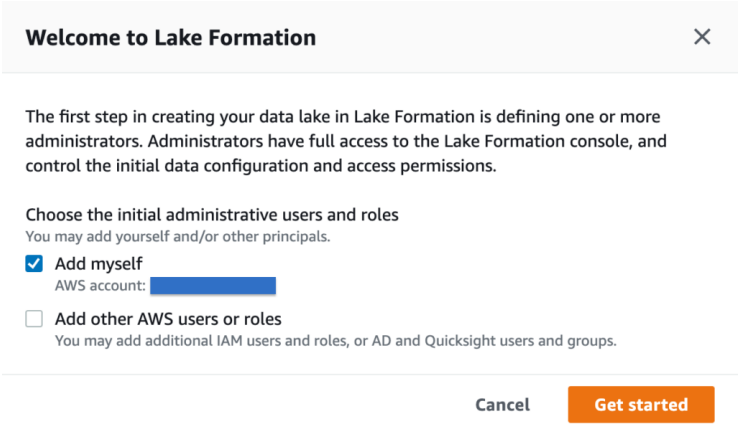
Prima di concentrarci sui vantaggi della gestione di un data lake tramite AWS Lake Formation, dobbiamo prima crearne uno in modo semplice.

Procediamo mediante console AWS e scegliamo "AWS Lake Formation" nell'elenco dei servizi o tramite la barra di ricerca. Ci verrà mostrata la seguente dashboard:



Schermata di accesso a Lake Formation

Dopo aver cliccato su "Get started", ci verrà chiesto di impostare un amministratore per il data lake; è possibile aggiungere utenti e ruoli AWS disponibili sull'account a cui si è connessi. Selezionane uno adatto, preferibilmente un ruolo, assumibile con credenziali temporanee da persone e servizi, e prosegui.



Selezionare un utente o un ruolo

Ottenuto l'accesso alla dashboard di Lake Formation, è il momento di aggiungere una ”Lake Location”, ovvero un percorso S3 valido da cui recuperare i dati. I dati possono essere ottenuti tramite vari sistemi, ad esempio mediante **Jobs di AWS Glue**, attraverso la combinazione di **AWS Kinesis stream** e **Data Firehose**, o semplicemente caricando i dati direttamente su **S3**.

Esaminiamo rapidamente tutte le possibilità per popolare il nostro **Glue Catalog** (strumento che definisce il nostro data lake dietro le quinte).

Innanzitutto, aggiungeremo la posizione del data lake facendo clic sul pulsante "Register location" nella sezione “Register and ingest” dalla dashboard del servizio, come in figura.

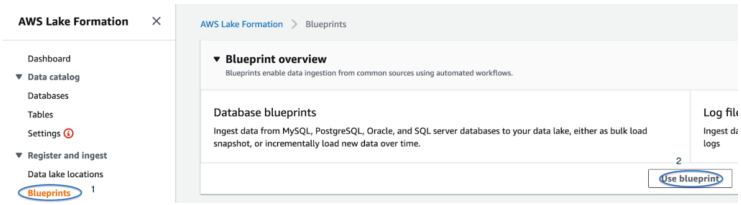
Aggiungiamo una nuova location per il data lake

Ci verrà richiesto di selezionare un bucket S3, procediamo, quindi aggiungiamo uno IAM role adatto (o permettiamo ad AWs di crearne uno), quindi completiamo la procedura con “Register location”.

Ora possiamo:

1. Caricare semplicemente i dati su S3 prima di iniziare un processo di crawling;
2. Utilizzare una combinazione di servizi AWS per fare l'ingestion dei dati, come Kinesis Stream e Firehose (leggete la nostra story per maggiori informazioni);
3. Utilizzare un **blueprint** di Lake Formation per ottenere velocemente dati da sorgenti di log o database di varia natura;

Analizziamo brevemente la terza opzione, potenzialmente limitata, ma molto interessante e non ancora coperta dai nostri blog post precedenti



Attivare un crawling job mediante un Blueprint

Clicchiamo su "Use blueprint", ci verrà presentato un form dove potremo selezionare se prelevare i dati da un database o una sorgente di log.

Ora ci basterà seguire le istruzioni per generare un workload, che in pratica è un ETL Job di Glue dove tutte le opzioni per le fasi di Extract, Transform, e Load si trovano in un unico punto.

Ad esempio, per un database MySQL, MSSQL o Oracle, aggiungiamo (o creiamo) una connessione AWS Glue, specificando anche il DB di origine e la tabella, secondo questo formato: **<db_name> / <table_name>**. Aggiungiamo (o creiamo) il Catalogo Glue di destinazione, specificando un DB e una tabella, utilizziamo anche anche lo strumento fornito, per selezionare un percorso S3 adatto ad ospitare i dati del catalogo.

Selezioniamo un nome per il flusso di lavoro, decidiamo la frequenza del crawler, ad esempio "Run on demand" ed infine, un prefisso per la tabella, le altre opzioni possono essere lasciate come predefinite.

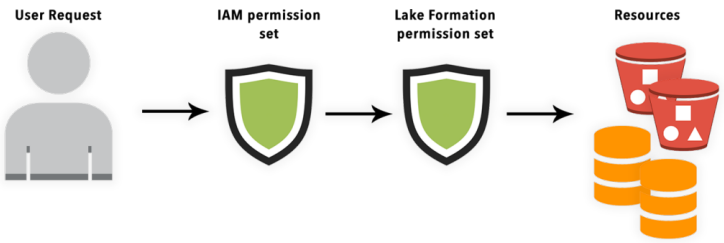
Alcune note: optare sempre per il formato **parquet** nella sezione target di S3, in quanto garantisce un solido incremento delle prestazioni sulle operazioni che verranno

eventualmente eseguite sul set di dati in seguito. Inoltre, se si prevede di utilizzare Athena per interrogare il proprio catalogo, utilizzare "_" invece di "-" per i nomi di database e tabelle, poiché quest'ultimo carattere a volte può portare a problemi di compatibilità indesiderati.

Sicurezza migliorata

Una volta che Lake Formation è pronto, possiamo focalizzarci sui dettagli che lo rendono uno strumento davvero degno di nota: in primis un modello a permessi “laschi” che va ad aumentare quello già fornito da IAM.

Un modello centralizzato così definito, permette di abilitare in seguito dei permessi molto più fini e granulari, mediante un semplice sistema di grant/revoke, così come mostrato in figura:



Come una richiesta passa attraverso 2 stage di permessi prima di avere accesso alle risorse

Le autorizzazioni di Lake Formation vengono applicate anche a livello di tabella e di colonna e funzionano su tutto lo stack completo di servizi AWS per l'analisi e l'apprendimento automatico, inclusi, ma non limitati a, Amazon Athena, Redshift, Sagemaker e vengono direttamente mappati su oggetti S3 in modo trasparente.

Il controllo degli accessi in AWS Lake Formation si divide in due macro aree distinte:

- **Metadata access control** – Autorizzazioni per le risorse del Glue Data Catalog che consentono ai **principal** di creare, leggere, aggiornare ed eliminare database e tabelle di metadati.
- **Underlying data access control** – Autorizzazioni per S3 che includono l'accesso ai dati e le autorizzazioni alle location (percorsi) dei dati. Le autorizzazioni di accesso ai dati consentono ai **principal** di leggere e scrivere oggetti su S3. Le autorizzazioni

per le location dei dati consentono la creazione di database e tabelle di metadati che puntano a path specifiche su S3.

L'utilizzo di questi elementi per centralizzare le policy di accesso ai dati è semplice: prima blocchiamo qualsiasi accesso diretto ai bucket richiesti su S3, in modo che Lake Formation gestisca tutti gli accessi ai dati.

Successivamente, configuriamo la protezione dei dati e le **policy di accesso** affinché vengano applicate a tutti i servizi AWS che accedono ai dati nel data lake. Sfruttando le autorizzazioni per metadati e oggetti, possiamo configurare utenti e ruoli per accedere solo a dati specifici fino a livello di tabella e/o colonna.

Prima di assegnare policy a utenti e risorse, Lake Formation necessita di alcune "personas", obbligatorie per funzionare correttamente, e queste sono anche necessarie per garantire la compatibilità con le versioni precedenti delle autorizzazioni gestite da IAM, create in precedenza per S3, AWS Glue, Athena e altri servizi:

IAM administrator

Utente che può creare altri utenti e ruoli IAM. Dispone della policy gestita da AWS "AdministratorAccess". Può anche essere designato come amministratore del data lake.

Data lake administrator

Utente che può registrare location su Amazon S3, accedere al catalogo dei dati, creare database, creare ed eseguire flussi di lavoro, concedere autorizzazioni a Lake Formation ad altri utenti e visualizzare i log di AWS CloudTrail.

Workflow role

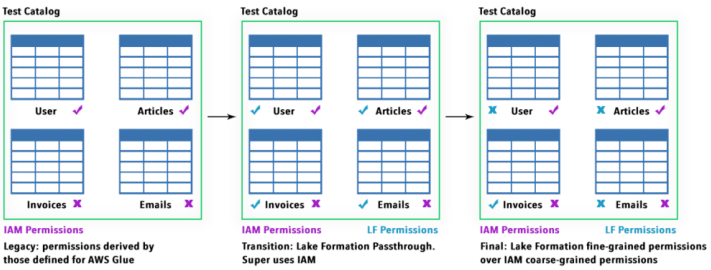
Ruolo che esegue un flusso di lavoro per conto di un utente. Si specifica questo ruolo quando un flusso di lavoro viene creato a partire da un progetto.

Le prime due "personas" sono anche definite **IAMAllowedPrincipals** e hanno attive le autorizzazioni "Super" e "Use only IAM access control" di default, garantendo così la retrocompatibilità con i workflow Glue preesistenti, S3 e Athena.

Come funzionano i permessi

Le autorizzazioni fine-grained sono organizzate in modo tale da permettere a Lake formation di sostituire le autorizzazioni IAM di tipo coarse. Questo al fine di garantire transizioni più fluide dal vecchio set di autorizzazioni a quello gestito interamente da Lake Formation.

Un semplice schema illustra le scelte possibili:



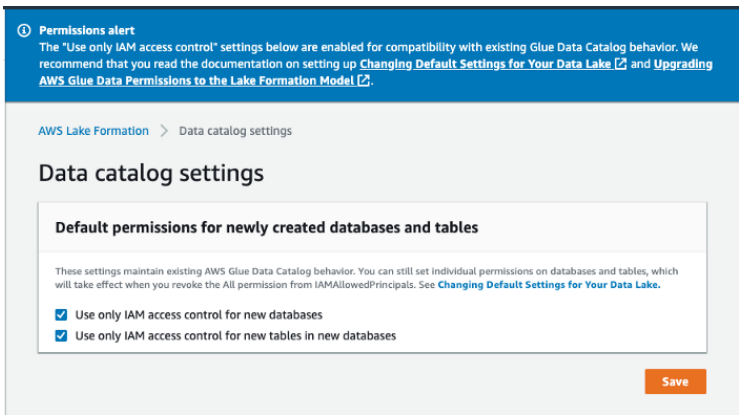
Come funziona la transizione del set di permessi da IAM a Lake Formation

Per vedere tutte le tipologie di permessi disponibili su Lake Formation, potete dare una lettura alla documentazione ufficiale di AWS.

Lake Formation attualmente supporta anche la Server-Side-Encryption su S3, nonché gli endpoint privati per le VPC.

Registra inoltre tutte le attività in AWS CloudTrail (che può anche essere un set di dati supportato), dando un ottimo contributo all'isolamento della rete e alla sua verificabilità.

I permessi di Lake Formation **si applicano solo nella Regione in cui sono stati concessi**.Per compatibilità con workflow precedenti, Lake Formation **passa attraverso le autorizzazioni IAM anche per le nuove risorse**, a meno di non **specificare un comportamento diverso**.



Notifica sull'uso di permessi legacy

Governance migliorata

Ciò che aiuta davvero a mantenere il controllo sul proprio data lake è che con Lake Formation, abbiamo finalmente una dashboard centralizzata per controllare le location su S3, i job ETL, i crawler, i cataloghi di GLue e ovviamente le autorizzazioni.

Un'altra caratteristica interessante è che Lake Formation viene fornito con Cloud Trail abilitato, quindi ogni azione eseguita dagli utenti o dai servizi, tramite i ruoli IAM, viene controllata e registrata direttamente nella dashboard.

Cloud Trail mostra le attività di accesso direttamente dalla dashboard di Lake Formation

Un'altra questione che dobbiamo gestire quando abbiamo a che fare con i data lake è la deduplicazione e la pulizia dei dati, che, se ignorate, portano a dati incoerenti, inefficienti e spesso inaccessibili.

Incapsulando le funzionalità di AWS Glue, Lake Formation offre **FindMatches ML Transform**: un Glue Job utilizzato per rimuovere i dati duplicati sfruttando gli algoritmi di Machine Learning. È possibile selezionare una soglia per la Accuracy, per indicare la quantità di precisione che l'algoritmo deve utilizzare per identificare dati potenzialmente duplicati (più precisione, più costi).

Per saperne di più, vi invitiamo a seguire questo tutorial su AWS.

Feature in Preview

Transazioni - inserire, eliminare e modificare righe contemporaneamente

Per gestire l'evoluzione dei dati nel tempo è fondamentale definire procedure in grado di mantenere il data lake sempre aggiornato. Questo è fondamentale in quanto l'accesso ai dati deve essere concesso a diversi utenti in qualsiasi momento e dobbiamo anche garantire l'integrità dei dati stessi. I dati sono anche spesso organizzati in modi strutturati e non strutturati contemporaneamente.

L'implementazione degli aggiornamenti in tempo reale è complessa e difficile da scalare. AWS Lake Formation introduce, in anteprima, nuove API che supportano le transazioni ACID utilizzando un nuovo tipo di tabella di data lake, chiamata Governed table.

Consente a più utenti di inserire, eliminare e modificare contemporaneamente righe tra tabelle, consentendo comunque ad altri utenti di eseguire query e modelli di machine learning sugli stessi set di dati, con la certezza che i dati siano sempre effettivi.

Sicurezza a livello di riga

Assicurarsi che gli utenti abbiano accesso solo ai dati corretti in un data lake è difficile. Gli amministratori di data lake spesso, mantengono più copie dei dati per applicare criteri di sicurezza diversi per utenti diversi. Ciò aggiunge complessità, sovraccarico operativo e costi di archiviazione aggiuntivi.

AWS Lake Formation consente già di impostare policy di accesso per nascondere i dati, anche per colonna, ad esempio i numeri di previdenza sociale.

Con la sicurezza a livello di riga, possiamo concedere autorizzazioni speciali per riga a utenti e ruoli, ovvero l'accesso a dati regionali specifici o dati relativi a un conto bancario specifico.

Accelerazione

Migliori prestazioni con filtri, aggregazioni e compattazione automatica dei file, grazie a un nuovo ottimizzatore di archiviazione che combina automaticamente file piccoli in file più grandi per velocizzare le query fino a 7 volte. Il processo viene eseguito in background, senza alcun impatto sulle prestazioni sui carichi di lavoro di produzione.

Referenze

https://pages.awscloud.com/Lake_Formation_Feature_Preview.html

https://aws.amazon.com/lake-formation/faqs/

https://docs.aws.amazon.com/lake-formation/latest/dg/access-control-overview.html

https://docs.aws.amazon.com/lake-formation/latest/dg/permissions-reference.html

Per riassumere

Nel corso dell'articolo abbiamo visto tutte le funzionalità che rendono Lake Formation una scelta adatta per la gestione dei data lake su AWS.

Ci siamo concentrati principalmente sugli aspetti di sicurezza e governance di questo servizio, mostrando come la gestione delle autorizzazioni a livello di oggetto su S3, possa essere un processo complesso, semplificato però dai permessi di Lake Formation.

Abbiamo mostrato come Lake Formation consenta di concedere / revocare autorizzazioni a utenti o ruoli, sia a livello di tabella che di colonna.

Abbiamo visto che le autorizzazioni di AWS Lake Formation sono più adatte delle autorizzazioni IAM per proteggere un data lake, perché vengono applicate su oggetti logici come un database, una tabella o una colonna invece che su file e directory; forniscono inoltre un controllo granulare per l'accesso a livello di colonna.

Abbiamo anche visto che queste autorizzazioni sono mappate internamente agli oggetti sottostanti che si trovano in S3.

Grazie a un'interfaccia utente semplice, non è necessario tenere aperte più schede per tenere traccia dei lavori ETL, delle posizioni S3 e dei cataloghi di dati per i nostri workflow. Tutte queste informazioni risiedono in un'unica dashboard, dove possiamo revocare o concedere direttamente i permessi degli oggetti che vi risiedono.

Abbiamo dato una rapida occhiata anche alle nuove funzionalità disponibili in preview. In particolare, un nuovo tipo di tabella, la tabella **governed**, che consente transazioni senza interruzioni per mantenere i dati sempre aggiornati. La possibilità di utilizzare una policy di accesso per riga e un nuovo ottimizzatore di archiviazione per aumentare le prestazioni nella gestione di grandi quantità di piccoli file.

Nonostante la gestione delle autorizzazioni, il workflow di importazione dei dati è semplificato, ma la maggior parte dei processi di Glue come ETL, Crawler, trasformazioni specifiche di ML devono essere impostate comunque manualmente.

Ed eccoci arrivati alla fine! Lasciateci un commento o contattateci per qualsiasi dubbio, domanda o idea!

Ci vediamo puntuali tra due settimane con un nuovo articolo su **#proud2becloud**!

Alessandro Gaggia

Head of software development di beSharp, Full-Stack developer, mi occupo di garantire lo stato dell'arte di tutta la nostra codebase. Scrivo codice in quasi ogni linguaggio, ma prediligo Typescript. Respiro Informatica, Game design, Cinema, Fumetti e buona cucina. Disegno per passione!

Copyright © 2011-2021 by beSharp srl - P.IVA IT02415160189