Democratizzare l'accesso ai dati tramite una Data Platform self-service – Parte 1

11 Febbraio 2025 - 9 min. read

Data Ingestion *Data Platform* *Medallion architecture*

In questa serie di articoli descriveremo come creare e strutturare correttamente una Data Platform self-service per la democratizzazione dei dati e l'analisi su AWS.

Partiremo dall'acquisizione e dall'archivio dei dati, passando per gli strumenti di elaborazione necessari a generare informazioni di valore per analisi, visualizzazioni e report. Inoltre, ci concentreremo su governance dei dati, reperibilità e collaborazione, con particolare attenzione alla sicurezza e al controllo degli accessi.

TL;DR

Lo standard *de facto* per le moderne piattaforme dati è l'architettura "a medaglione". Tutte le sorgenti dati, sia in streaming che in batch, vengono acquisite nel cosiddetto "bronze layer". L'acquisizione dei dati può essere effettuata con strumenti come AWS DMS e AWS Glue. I dati grezzi devono poi essere puliti, normalizzati e strutturati prima di essere archiviati nella "silver area". Ulteriori trasformazioni organizzano i dati per casi d'uso aziendali (AI, reporting, ecc.) nel "golden layer".

I "Golden data" sono poi pronti per essere utilizzati dagli utenti aziendali.

Perché una Data Platform?

Nell'era moderna, c'è una frase che continua a riecheggiare: "I dati sono il nuovo oro". Ma è davvero così? Spoiler: SÌ!

Ogni clic, acquisto o interazione lascia tracce di informazioni preziose che, se analizzate correttamente, possono trasformare il modo in cui le aziende operano. Sfruttando il potere dei dati, le imprese possono comprendere meglio il proprio mercato, individuare schemi di comportamento dei clienti e migliorare l'efficienza operativa. Integrare le decisioni basate sui dati nei processi decisionali aziendali affina le strategie, consentendo di fare scelte informate basate su dati reali, migliorando così produttività e performance.

Avere dati è sicuramente importante, ma se non puoi utilizzarli, averli diventa quasi inutile!

Come fare, quindi? Ecco che si fa strada un altro concetto sempre più popolare: Data Platform!

Una Data Platform è una soluzione composta da diversi elementi infrastrutturali che raccoglie tutti i dati dell'organizzazione nella loro forma più grezza e, attraverso più fasi, li trasforma fino a restituirli nella loro forma ottimale. Proprio come il taglio e la lucidatura delle gemme, la Data Platform estrae il massimo valore dai dati non strutturati e grezzi.

Per le aziende, una Data Platform è un asset strategico che abilita le decisioni basate sui dati, alimenta l'innovazione e fornisce un vantaggio competitivo nell'attuale economia digitale.

L'architettura a medaglione

L'architettura più comune per una Data Platform è l'**architettura a medaglione**, che suddivide il processo di raffinazione dei dati in tre livelli: bronze, silver e gold.

L'idea centrale è che il **bronze layer** sia l'area di atterraggio per i dati nella loro forma più grezza. Tutti i processi di raccolta trasportano i dati da fonti esterne, come database e API, al livello bronzo, solitamente aggiungendo campi per la tracciabilità e il controllo. Le strategie di lettura più comuni includono Change Data Capture (CDC), caricamenti batch incrementali e, in alcuni casi, caricamenti completi delle sorgenti. Oltre ai processi di raccolta, il livello bronzo può anche fungere da area di atterraggio per sistemi che inviano direttamente i dati, come file CSV/Excel o dati forniti manualmente, sebbene possa essere utile un'area di "quarantena" per i dati provenienti da sistemi non affidabili.

Una volta raccolti tutti i dati dalle diverse fonti, è il momento di elaborarli. I processi di trasformazione che trasportano i dati dal livello bronzo si occupano della pulizia e della validazione per garantire un certo livello di qualità. Qui troviamo attività come strategie di deduplicazione, gestione dei valori nulli e trattamento degli errori. Inoltre, i dati vengono strutturati secondo schemi standard prima di passare al livello successivo: il “silver layer”.

Il **silver layer** dell'architettura a medaglione contiene dati puliti e strutturati. Questo livello ospita una forma primordiale di dati interrogabili, pronti per essere consumati. Gli utenti aziendali possono iniziare ad esplorare questi dati e creare nuovi casi d’uso. È fondamentale strutturare correttamente i dati all’interno della Data Platform affinché siano facilmente accessibili, promuovendo così la democratizzazione dei dati e l’analisi self-service.

Prima di arrivare all’ultimo livello, i dati raffinati nel livello argento vengono ulteriormente trasformati e aggregati per rispondere alle esigenze aziendali e ai casi d’uso definiti dagli utenti di business. L’obiettivo principale di questi processi di trasformazione è strutturare i dati per ottimizzare le performance delle query e rispondere rapidamente alle richieste aziendali.Il **gold layer** contiene i dati nella loro forma più raffinata e di massimo valore. I dati in questo livello sono ottimizzati per servire come fonte per utenti aziendali come Business Analyst, Data Engineer e Data Scientist, supportando casi d’uso come business intelligence, report finanziari, AI e machine learning.

L’acquisizione del dato: il setup

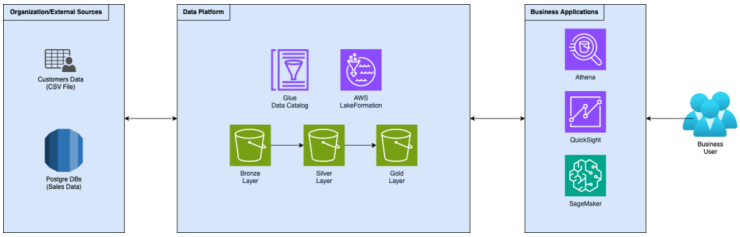
Introduciamo brevemente il caso d’uso di questo articolo per avere una panoramica generale di ciò che vedremo, riprendendo i concetti discussi in precedenza per comprendere meglio la soluzione.

Immaginiamo di essere un'azienda fittizia che aiuta altre aziende e liberi professionisti, operanti in diversi settori, a migliorare le loro performance di vendita e, di conseguenza, ad aumentare i loro ricavi. L'azienda vuole iniziare a sfruttare i dati dei clienti per estrarre informazioni di valore, migliorare le strategie proposte e supportarli meglio nell'incremento delle loro entrate. In particolare, il business ritiene che raccogliere dati ed eseguire analisi di Business Intelligence possa essere un primo passo efficace!

Tu, come parte del Data Team, hai il compito di creare l'intera infrastruttura a supporto di questi casi d'uso, insieme alla logica del codice necessario.

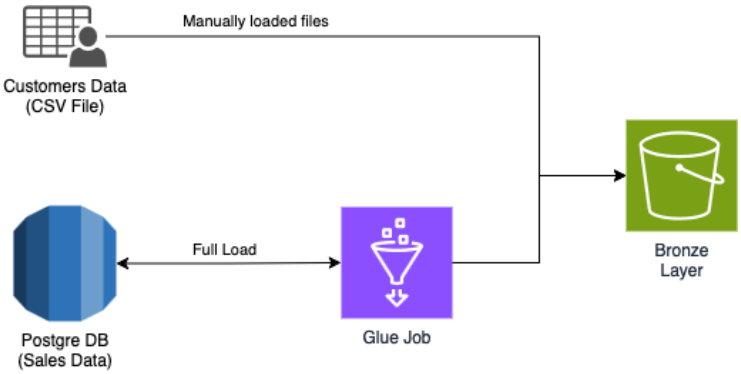
Per completare la configurazione, supponiamo che la tua azienda abbia due clienti appartenenti a due settori diversi: elettronica e alimentare. Hai a disposizione i loro dati di vendita, insieme alle informazioni sui rispettivi clienti, provenienti da diverse fonti: database relazionali e file CSV.

Ecco un diagramma ad alto livello del setup architetturale:



Come puoi vedere, e come abbiamo descritto in precedenza, i dati fluiscono nella Data Platform e vengono raffinati iterativamente attraverso i livelli dell'architettura a medaglione. All'interno del livello oro, i dati sono progettati per essere consumati da applicazioni aziendali come AI (SageMaker), BI (QuickSight) o CI (gioco di parole intenzionale!).

Poiché stiamo creando una Data Platform semplice con dati di esempio, abbiamo scelto il metodo più rapido possibile: i file dei dati dei clienti sono stati caricati manualmente nel bucket del livello bronzo, mentre per i dati di vendita abbiamo creato un Glue Job, insieme a una Glue Connection, per gestirne il caricamento.



Ovviamente, nella creazione di una vera Data Platform, il contesto sarà molto diverso da questo esempio semplificato e le scelte a disposizione saranno molteplici. Ecco alcuni consigli per gestire diversi tipi di dati:

I Glue Job offrono la funzionalità Bookmark, che permette di leggere una sorgente dati riprendendo dall'ultimo punto in cui si era interrotto. Per quanto riguarda i database, è

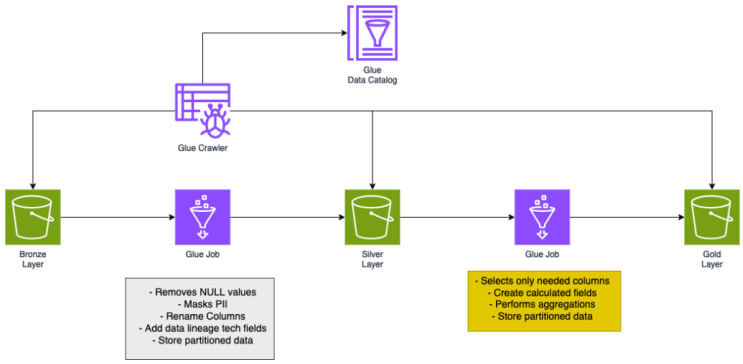
possibile utilizzare AWS DMS con la strategia CDC (Change Data Capture) per acquisire le modifiche in tempo reale.

Per le API, si possono sfruttare le Lambda Functions per orchestrare l'ingestione dei dati.Per i dati IoT, AWS offre le suite **IoT Core** e **IoT Analytics**, che includono un broker MQTT per acquisire i dati dai dispositivi e archiviarli direttamente su S3, con la possibilità aggiuntiva di trasformare i dati già in fase di acquisizione.

Data Pipeline: dall'acquisizione ai "golden data"

Ora che abbiamo descritto alcune strategie di acquisizione dei dati, immergiamoci nel mondo della data pipeline: dal livello bronzo fino ai dati oro, pronti per essere consumati.

Ecco un'immagine che illustra la struttura della data pipeline.



Come puoi vedere, sfruttiamo la potenza dei Glue Job, e quindi di un cluster Spark, per far fluire i dati all'interno della pipeline. Inoltre, utilizziamo il Glue Data Catalog insieme ai Glue Crawlers per rendere i dati interrogabili.

Da Bronze a Silver

Come abbiamo descritto poco fa, le trasformazioni che permettono di portare i dati dal livello bronze al livello silver si concentrano principalmente sulla pulizia del dato e la standardizzazione.

Le procedure di pulizia del dato possono variare molto a seconda del caso d'uso, alcune operazioni comuni che vengono fatte riguardano la gestione dei dati mancanti, l'eliminazione di dati duplicati e la gestione degli outliers. Nel nostro caso dobbiamo semplicemente gestire i dati mancanti, essendoci poche righe con dati mancanti abbiamo deciso di cancellarle direttamente.

Infine, abbiamo mascherato i dati sensibili prima di procedere con gli step successivi.

In accoppiata alla pulizia dei dati è molto comune trovare controlli sulla qualità stessa dei dati. Un'operazione comune da effettuare è la validazione delle colonne che devono avere una particolare struttura, per esempio il campo "email", o che devono contenere un numero in un particolare range di valori possibili.

Dopo aver pulito i dati è il momento di standardizzare lo schema. Per farlo abbiamo rinominato alcune colonne e aggiunto alcuni campi tecnici per aiutarci a tenere traccia del data lineage, i campi in questione sono l'id del lavoro e l'orario corrente.

Finalmente siamo pronti per memorizzare i dati trasformati dentro al livello silver. Abbiamo salvato i dati in formato parquet, un formato molto compresso ed efficiente. Per i dati riguardanti le vendite abbiamo partizionato per negozio e prodotto così che le query di lettura possano essere ottimizzate.

Da Silver a Gold

Le trasformazioni per passare dal livello silver al livello gold sono legate alle necessità del business.

Nel nostro esempio, il business voleva semplicemente dei grafici dei ricavi totali in un determinato periodo di tempo.

Per portare a termine questo compito, siamo partiti selezionando solamente i campi necessari al calcolo dei ricavi totali, eliminando tutto ciò che è superfluo.

Abbiamo calcolato i ricavi moltiplicando il prezzo e la quantità venduta, così facendo abbiamo creato un nuovo campo, solitamente vengono chiamati "campi calcolati".

Successivamente, abbiamo aggregato rispetto ai dati calcolati per ottenere i ricavi totali. Infine, abbiamo aggregato per data e negozio per poter fare un'operazione di somma finale.

Finalmente, i dati sono pronti per essere memorizzati nel livello gold. Abbiamo memorizzato i dati in formato parquet, partizionando per negozio e data.

Ora il business può capire meglio come aiutare i propri clienti a incrementare le vendite.

Punti principali e piccoli spoiler

In questo articolo hai visto un esempio di una data platform, partendo dal perché è utile, visionando le architetture più comuni, l'architettura a medaglione, e le varie trasformazioni effettuate sui dati, dalla loro forma più grezza fino al dato raffinato che porta maggior valore.

Nei prossimi capitoli di questa serie porteremo il livello ancora più in alto, spiegando soluzioni per poter migliorare la data platform portandola ad essere una self-service platform, promuovendo così la democratizzazione dei dati. Inoltre, useremo AWS LakeFormation per gestire le utenze ed i permessi di accesso ai dati.

Hai mai provato a creare una data platform per conto tuo?

Raccontaci la tua esperienza nei commenti. A presto con la seconda parte!

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.
