Democratizzare l'accesso ai dati tramite una Data Platform self-service utilizzando AWS LakeFormation – Parte 3

20 Maggio 2025 - 9 min. read

Amazon SageMaker | Data Ingestion | Data Platform | Machine Learning | Medallion architecture | ML

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. Abbiamo iniziato con l'acquisizione e l'archivio dei dati per poi passare agli strumenti di elaborazione per creare dati preziosi per analisi, visualizzazioni e reportistica. Inoltre, ci siamo contrati sulla governance dei dati, sulla reperibilità e sulla collaborazione, con un occhio alla sicurezza e al controllo degli accessi.

Questo articolo conclude questa serie iniziata con la descrizione delle piattaforme dati e delle relative pipeline di dati. Poi, ci siamo soffermati sulla governance dei dati e abbiamo approfondito la democratizzazione dell'accesso ai dati attraverso una data platform self-service, utilizzando AWS LakeFormation.

Vedremo come estrarre il vero valore dai tuoi dati utilizzando SageMaker per creare un modello di Machine Learning per prevedere i dati di vendita e QuickSight per creare visualizzazioni che mostrino come il modello prevede i dati futuri.

TL;DR

Estrai il massimo valore dai dati costruendo applicazioni su di essi, come modelli di Machine Learning (ML) per le previsioni o report di Business Intelligence (BI) per

visualizzare le tendenze. Utilizza le funzionalità di SageMaker Unified Studio Experience per creare il modello ML. Esegui l'Analisi Esplorativa dei Dati (EDA) utilizzando notebook, addestra diversi modelli con pipeline (visuali) e seleziona il migliore dal registro dei modelli. Crea dashboard e report utilizzando AWS QuickSight per mostrare le previsioni del modello, insieme ad altre metriche e KPI.

Estrarre valore dai dati

Negli articoli precedenti di questa serie, abbiamo imparato cosa sia una moderna data platform, come strutturarla correttamente con l'**architettura a medaglione** e implementarla su AWS. Con la data platform come fondamento della nostra architettura dati, ci siamo poi concentrati sull'applicazione della governance dei dati e sulla democratizzazione dei dati nel secondo capitolo, utilizzando AWS LakeFormation. Stabilendo un data lake ben architettato con AWS LakeFormation, abbiamo visto come le aziende possono abbattere i tradizionali silos di dati mantenendo adeguati controlli di sicurezza e governance. Abbiamo esaminato come le organizzazioni possono trasformare le risorse di dati grezzi in risorse accessibili e gestite che danno potere ai team di tutta l'azienda. Questa democratizzazione dei dati crea la base essenziale per ciò che segue nella catena del valore dei dati.

In questo terzo capitolo, sposteremo il nostro focus dall'infrastruttura dati alle applicazioni di dati, gli strumenti e i sistemi potenti che trasformano i dati strutturati in informazioni e azioni. Nello specifico, esploreremo come i modelli di machine learning e le tecniche di visualizzazione dei dati possono essere implementati su AWS per estrarre il massimo valore dalla tua data platform.

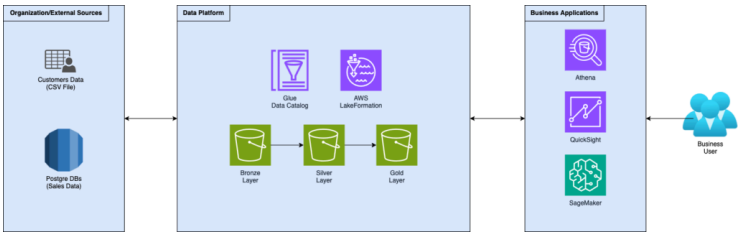
Queste applicazioni rappresentano l'ultima tappa del nostro percorso dati: i dati, adeguatamente organizzati e accessibili diventano il carburante per l'analisi predittiva e i processi decisionali che guidano i risultati aziendali.

Dalla Data Platform alle Applicazioni di Dati

Se hai letto il primo di questa serie di articoli, già sai con cosa stiamo lavorando ma, per essere tutti allineati, ecco una breve panoramica del setup.

Agendo come ingegneri dei dati per un'azienda fittizia che aiuta i suoi clienti ad aumentare i loro ricavi, abbiamo creato una data platform seguendo la, ormai standard, architettura a medaglione. Abbiamo sviluppato logiche di acquisizione e trasformazione per raccogliere dati e spostarli attraverso i livelli sempre più raffinati

della data platform. Quindi, abbiamo implementato la governance utilizzando AWS LakeFormation, rendendo i dati accessibili ai team interni e ai clienti.



È passato del tempo da quando tuo cliente ha adottato la strategia proposta per aumentare le vendite ed è piuttosto soddisfatto, ma si chiede: "fino a che punto possiamo spingerci con questa strategia?". L'azienda ora ti chiede di iniziare a estrarre valore reale dai dati.

L'idea è di creare un modello che, prendendo i dati di vendita dei mesi passati, preveda le vendite per il prossimo anno, mese per mese. Puntiamo anche a mostrare al cliente varie intuizioni per aiutare la creazione della strategia di vendita per il prossimo periodo mostrando loro la linea di tendenza delle vendite previste, insieme ai loro prodotti più venduti e meno venduti, in modo che possano decidere quali prodotti vale la pena vendere e quali possono essere eliminati dal loro catalogo.

Amazon SageMaker Unified Studio Experience

Ora che abbiamo fissato gli obiettivi, introduciamo il nostro primo strumento.

Per i più attenti tra voi, all'interno dell'ultimo capitolo, abbiamo scritto di Amazon DataZone, uno strumento che si basa su AWS LakeFormation per gestire facilmente la governance e la condivisione dei dati. Tuttavia, l'intero ecosistema AWS AI è in fase di riprogettazione e rebranding, fortunatamente per noi!

Tutto ora ricade sotto la pagina di destinazione di SageMaker che ci indirizza attraverso i vari servizi AI, dai data warehouse e motori di query (Athena e Redshift), alle trasformazioni dei dati (EMR e Glue), alla modellazione (SageMaker AI, ex SageMaker), alla AI generativa (Bedrock) e alla governance dei dati (LakeFormation e DataZone).

Di conseguenza, anche le varie esperienze "Studio" sono state aggregate sotto un'unica suite completa chiamata **Amazon SageMaker Unified Studio**. Questa nuova esperienza Studio dovrebbe coprire l'intero percorso dei dati dalle fondamenta fino al prodotto finale: l'applicazione AI. Dalle basi dell'elaborazione dei dati, allo sviluppo del

modello o all'AI generativa, fino alla distribuzione su larga scala, supportando il lavoro con notebook e editor SQL integrato. Il SageMaker Unified Studio ha le capacità di DataZone, quindi puoi organizzare risorse e utenti in "domini" e promuovere la collaborazione attraverso "progetti".

Utilizzeremo il SageMaker Unified Studio per creare il nostro modello ML.

Previsione delle vendite

Utilizzando Amazon SageMaker Unified Studio, iniziamo con la creazione della nostra prima applicazione dati: il modello di Machine Learning.

Per costruire un modello di machine learning è essenziale avere dati e, soprattutto, conoscere e comprendere i tuoi dati. Iniziamo raccogliendo i dati ed eseguendo alcune analisi standard.

Analisi Esplorativa dei Dati (EDA)

Abbiamo dati, già puliti e preparati, all'interno del nostro gold layer della data platform. Carichiamoli e iniziamo a fare alcune analisi su di essi.

Una delle feature di SageMaker Unified Studio sono le istanze notebook, ne useremo una come nostra unità di elaborazione per esplorare i nostri dati. Ecco come appare l'esperienza notebook all'interno dello studio:

Come potresti ricordare dagli episodi precedenti di questa serie di articoli, stiamo utilizzando dati di esempio, appositamente creati a supporto di esso. Questo ci aiuta poiché durante il processo di creazione, abbiamo specificato le proprietà statistiche dei dati.

Senza ulteriori spiegazioni, vediamo cosa emerge dalla nostra EDA:

- 7 anni di dati di vendita: da gennaio 2018 a dicembre 2024
- I dati di vendita hanno un chiaro trend: aumento del 30% ogni anno

- I dati di vendita hanno stagionalità: dal 10% fino al 40% di aumento durante i mesi estivi e invernali

Ecco un grafico che rappresenta la descrizione dei dati:

Trovare il miglior modello

Ora abbiamo una buona comprensione dei nostri dati e delle relative proprietà statistiche. Questo tipo di conoscenza è cruciale per questa prossima sezione: è il momento della modellazione!

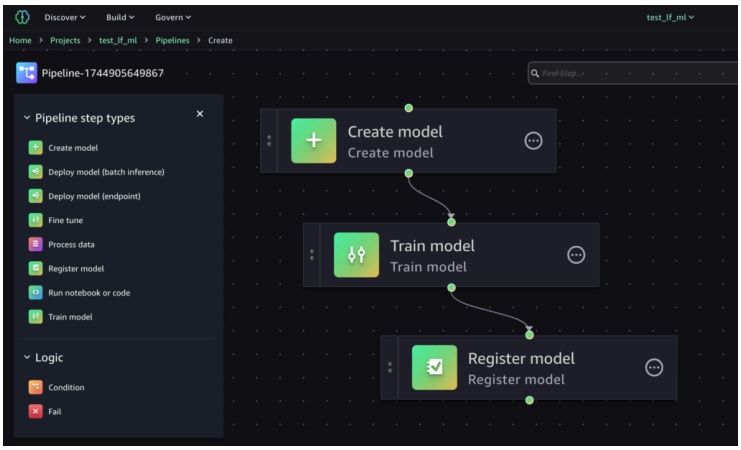
Stiamo lavorando con dati di serie temporali, mirando a prevedere le vendite per i mesi futuri, quindi, cerchiamo di trovare un buon modello da una semplice selezione di ciò che potrebbe essere adatto per questo caso d'uso:

- Holt-Winters Exponential Smoothing
- SARIMA
- Prophet

Per trovare il miglior modello all'interno di questo gruppo, dobbiamo definire un set di test.

Abbiamo 7 anni di dati, quindi possiamo fare training dei nostri modelli sull'intero dataset, separando l'ultimo anno, che possiamo utilizzare come nostro test set.

Possiamo addestrare i nostri modelli con un'altra feature di SageMaker Unified Studio: le Pipeline.



Possiamo creare l'intera pipeline ML con pochi click, utilizzando l'editor visuale.

Come si può vedere dall'immagine, ci sono svariate possibilità nella definizione della Pipeline ML che permettono di fare tutte le classiche attività necessarie per questo tipo di lavoro.

Nel nostro esempio, abbiamo bisogno di pochi semplici passaggi: definire il modello, addestrarlo e registrarlo all'interno del registro dei modelli, così da trovare le sue metriche e verificare la qualità del dato modello.

Dopo aver addestrato tutti i nostri modelli, scopriamo che il miglior modello è: Prophet!

A onor del vero, tutti i modelli sono molto vicini tra loro in termini di qualità poiché sono addestrati su dati appositamente creati per questo esempio.

Ora che abbiamo il nostro miglior modello, possiamo utilizzarlo per prevedere le vendite per i prossimi mesi e utilizzarlo nella nostra prossima sezione: visualizzazione dei dati e business intelligence.

Dalla Previsione alla Visualizzazione: BI in Azione

Ora che abbiamo le nostre previsioni di vendita per i prossimi mesi, integriamole con la BI.

La Business Intelligence (BI) mira a trasformare i dati in informazioni pratiche attraverso dashboard e report interattivi, monitorando KPI e visualizzando tendenze, permettendo un migliore processo decisionale.

Ritornando alla nostra narrazione, puoi mostrare al tuo cliente fittizio varie informazioni, come le potenziali prestazioni della strategia di vendita che hai proposto

durante il prossimo anno. Inoltre, potresti voler mostrare al tuo cliente i loro prodotti più venduti, insieme ai prodotti meno venduti, in modo che possano decidere quali prodotti vale la pena vendere e quali possono essere eliminati dal catalogo.

Creiamo questa dashboard con AWS QuickSight.

Prima di tutto, dobbiamo caricare i nostri dati importandoli nella sezione Datasets.

Per proiettare le vendite future abbiamo creato un file con le previsioni, contenente anche i dati storici, che possiamo importare direttamente.

Per quanto riguarda i prodotti più venduti e meno venduti, abbiamo questi dati all'interno del nostro silver layer della data platform. Importiamo questi dati utilizzando una connessione Athena con una query ad-hoc:

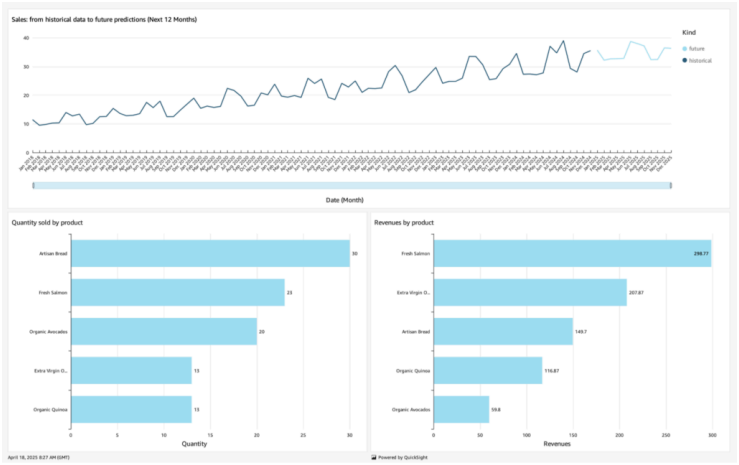
```
SELECT product, sum(quantity) as qta_sold, sum(quantity*price) as rev
enues
FROM "sales-processed-db"."lf_food_processed"
GROUP BY product;
```

Ora che abbiamo caricato i nostri dati in AWS QuickSight, creiamo un'Analisi per provare alcune visualizzazioni.

Possiamo utilizzare un grafico a linee per tracciare i dati delle vendite future, differenziando i dati storici dalle nostre previsioni.

I grafici a barre possono essere utilizzati per mostrare come i prodotti stanno vendendo in termini di quantità e ricavi.

Una volta che siamo soddisfatti del risultato, possiamo pubblicare la nostra Analisi come Dashboard in modo che gli utenti possano iniziare a beneficiarne. Ecco come appare con i nostri dati:



Conclusioni

Con questo articolo concludiamo il nostro viaggio nel mondo dei dati: dalla creazione di una data platform come repository centrale dei dati, passando dalla democratizzazione dei dati e dalla governance dei dati utilizzando AWS LakeFormation, all'estrazione di reale valore dai dati creando applicazioni su di essi.

Abbiamo esplorato la SageMaker Unified Studio Experience, eseguendo analisi esplorativa dei dati (EDA) utilizzando notebook e creando un modello AI in grado di prevedere dati futuri basandosi su serie temporali, utilizzando Pipeline e il Registro dei Modelli.

Infine, valorizziamo le previsioni del nostro modello, insieme ad ulteriori informazioni potenzialmente utili ai processi decisionali, mostrandole con una dashboard creata con AWS QuickSight.

Speriamo che abbiate apprezzato questo viaggio. Condividete pensieri, opinioni e sensazioni nella sezione 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.
