DATA ANALYTICS SU AWS: LA NOSTRA GUIDA INTRODUTTIVA

beSharp | 13 Novembre 2020

Con la società che evolve in una comunità completamente connessa digitalmente, la quantità di dati generati e raccolti sta crescendo notevolmente. Al giorno d'oggi, disponiamo di enormi quantità di dati sia sui processi aziendali che sui comportamenti dei clienti.

Questa situazione ci offre l'opportunità di sfruttare quantità sempre maggiori di dati di qualità sempre maggiore ad un costo ridotto.

Un esempio calzante è la profilazione del comportamento dei clienti: le aziende possono raccogliere informazioni su quali prodotti utilizzano i clienti, come li usano, quali aspetti di ogni prodotto sono realmente rilevanti nella loro vita quotidiana e molto altro ancora.

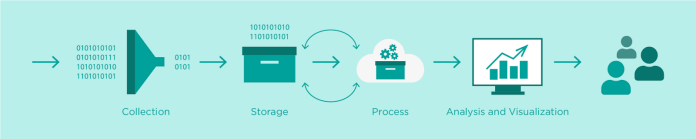
Passando dalla quotidianità all'industria, è oggi possibile ottenere informazioni su quali componenti dei macchinari sono soggetti ad usura e agire di conseguenza (manutenzione predittiva). È anche possibile ottenere dati sui pezzi difettosi per migliorare le linee di produzione, e così via.

La capacità di capire quali dati estrarre, di raccoglierli in modo efficiente, di archiviarli a basso costo e infine di analizzarli è quindi ciò che fa davvero la differenza, e rappresenta un notevole vantaggio competitivo.

L'effort richiesto per analizzare questa sorprendente quantità di dati risulta estremamente challenging utilizzando solo soluzioni tradizionali. AWS fornisce un'ampia gamma di servizi completamente gestiti creare applicazioni scalabili di data analytics in Cloud. Sia che la tua applicazione richieda lo streaming in tempo reale o l'elaborazione in batch, Amazon Web Services dispone dei servizi necessari per creare una soluzione completa di analisi dei dati.

La pipeline di Data Analytics

Per raccogliere, archiviare e analizzare i dati abbiamo bisogno di approfondire 4 aree che sono comunemente necessarie, indipendentemente dal tipo di progetto e implementazione: Collection, Data lake / Storage, Processing, Analysis and Visualization.



Quindi, in generale, i passaggi di una tipica pipeline di analisi dei dati possono essere riassunti come segue:

1. Un'infrastruttura appropriata raccoglie (ingestion) i dati dal campo.
2. I dati vengono archiviati utilizzando un servizio di archiviazione appropriato, ottimizzando il pattern di accesso ai dati.
3. I dati vengono quindi elaborati leggendo l'input dal servizio di archiviazione, eseguendo le operazioni richieste e quindi archiviando i dati elaborati in un'altra posizione.
4. Alla fine, le informazioni elaborate possono essere visualizzate utilizzando uno strumento di business intelligence (BI) per ottenere il valore per gli utenti finali e il business.

Nei capitoli seguenti, discuteremo ciascuno dei passaggi, con un focus sui servizi AWS che puoi sfruttare per implementare senza problemi i passaggi della pipeline.

Collection

Per progettare la corretta infrastruttura di ingestion, è necessario pensare alle caratteristiche dei dati e considerare le aspettative relative alla latenza, al costo e alla durevolezza dei dati.

Il primo aspetto da considerare è la frequenza di input, che è la misura della rapidità con cui i dati verranno inviati al sistema di raccolta. Ci si può riferire a questo KPI anche con il nome informale di "temperatura" (Hot, Warm, Cold) dei dati. La frequenza dei dati di input determina il tipo di infrastruttura da progettare. I dati transazionali (SQL) vengono meglio acquisiti utilizzando strumenti come **AWS Database Migration Service**, mentre i flussi di dati in tempo reale e near real-time sono il caso d'uso perfetto per **Amazon Kinesis Data Streams** e **Kinesis Firehose**.

Amazon Kinesis Data Streams è un modo affidabile, duraturo ed economico per raccogliere grandi quantità di dati dal campo e da applicazioni mobili o web. Uno dei molti vantaggi di Kinesis Data Streams è che è possibile estenderne le funzionalità con software personalizzato per soddisfare esattamente le esigenze di business. Può archiviare i dati raccolti per un massimo di 7 giorni e supporta molteplici applicazioni per lo stesso stream. Il software personalizzato può essere sviluppato sfruttando le **funzioni Lambda**.

Amazon Kinesis Data Firehose gestisce completamente molti dei processi manuali richiesti da Kinesis Data Streams e include anche opzioni di configurazione senza codice per fornire automaticamente i dati ad altri servizi AWS. Semplifica il raggruppamento dei dati in batch e la creazione di aggregazioni. I flussi di Kinesis Data Firehose possono essere configurati per inviare i dati aggregati verso **Kinesis Data Streams**, **Amazon Amazon S3**, **Amazon Redshift** e **Amazon ElasticSearch**.

I dati freddi generati da applicazioni, ovvero che possono essere elaborate periodicamente in batch, possono essere raccolti in modo efficiente utilizzando **Amazon EMR** o **AWS Glue**.

Un'altra caratteristica fondamentale da tenere in considerazione è il volume di dati. La quantità di dati da trasferire è un buon indicatore dei servizi che possono essere utilizzati. Alcuni servizi, come **Kinesis**, **SQS** e molti altri, hanno limiti sulle dimensioni dei record o elementi in ingresso. Pertanto, conoscere la "dimensione del blocco" dei dati di input è essenziale per progettare correttamente l'infrastruttura di importazione.

Altri servizi AWS hanno limiti sulla dimensione totale dei dati memorizzabili. Inoltre, durante la fase di progettazione è necessario tenere conto del throughput complessivo dell'importazione dimensionare correttamente la capacità di computing e lo stack di networking.

Storage / Data Lake

In questo step è necessario scegliere tra data warehouse o data lake.

Un data warehouse è un database ottimizzato per analizzare grandi quantità di dati relazionali provenienti da sistemi transazionali e applicazioni aziendali. La struttura e lo schema dei dati vengono definiti in anticipo ed è possibile ottimizzarli per query SQL veloci, in cui i risultati vengono generalmente utilizzati per l'analisi e il reporting operativo. I dati vengono puliti, arricchiti e trasformati in modo che possano fungere da repository centrale.

La scelta standard per creare un data warehouse su AWS è **Amazon Redshift**.

Con Redshift, è possibile eseguire query su petabyte di dati strutturati e semi strutturati utilizzando SQL standard. Redshift consente di salvare facilmente i risultati delle tue query su Amazon S3 utilizzando formati aperti come Apache Parquet per analizzarli ulteriormente da altri servizi di analisi come **Amazon EMR**, **Amazon Athena**e **Amazon SageMaker**.

Un data lake è un repository centralizzato in cui è possibile archiviare tutti i dati strutturati e non strutturati, indipendentemente dalla fonte o dal formato. Può memorizzare sia dati relazionali provenienti da applicazioni aziendali, sia dati non relazionali. La struttura dei dati (o schema) non è definita al momento dell'acquisizione. Ciò significa che è possibile memorizzare i dati senza sapere conoscerne la forma nè gli scopi per cui saranno impiegati. È inoltre possibile interrogare i data lake mediante svariati tipi di analisi sui dati, come query SQL, ricerca full-text, analisi in tempo reale e apprendimento automatico.

Per creare un data lake su AWS la scelta sconatata è **Amazon S3.**

Amazon Simple Storage Service (Amazon S3) è un servizio di storage ad oggetti che offre alta scalabilità, alta disponibilità dei dati, sicurezza e prestazioni. Può essere utilizzato con Amazon Athena e si integra anche con la maggior parte dei servizi utilizzati per creare pipeline di analisi dei dati.

Amazon Athena è un servizio di query interattivo che semplifica l'analisi dei dati in Amazon S3 utilizzando SQL standard. Athena è serverless, quindi non c'è infrastruttura da gestire e paghi solo per le query che esegui. La maggior parte dei risultati viene restituita in pochi secondi. Non sono

necessari lavori ETL complessi per preparare i dati per l'analisi. Ciò consente a chiunque abbia competenze SQL di analizzare rapidamente set di dati su larga scala.

Athena è integrato immediatamente con AWS Glue Data Catalog, consentendo di creare un repository di metadati unificato su vari servizi, eseguire la scansione delle origini dati per scoprire schemi e popolare il tuo catalogo con definizioni di partizioni e tabelle nuove e modificate e mantenerlo controllo delle versioni dello schema.

Processing

I dati grezzi sono raramente utili nell'analisi finale. È fondamentale preparare con cura i dati per aiutare gli analisti a trovare le informazioni di cui hanno bisogno. Il processo viene comunemente chiamato data wrangling. La preparazione dei dati permette l'aggiunta di campi calcolati, l'applicazione di filtri e la modifica dei nomi dei campi o dei tipi di dati.

La preparazione dei dati è un processo critico ed impegnativo. Per preparare i dati per l'analisi vanno prima estratti da varie fonti, quindi vanno ripuliti, trasformati nel formato richiesto e caricati in database, data warehouse o data lake per le analisi successive.

In AWS è possibile eseguire queste attività utilizzando i seguenti servizi: **Amazon Kinesis Data Analytics, Pre-elaborazione di Amazon Sagemaker**, **Amazon EMR** e **AWS Glue**.

Se si desidera analizzare i dati in streaming, sia in tempo reale che in near real-time, è possibile sfruttare direttamente Kinesis. **Amazon Kinesis Data Analytics**, è un ottimo strumento per realizzare trasformazioni i di base sui dati in streaming utilizzando i comandi SQL.

Amazon Sagemaker Preprocessing consente di avviare facilmente istanze EC2 su richiesta per eseguire processi di trasformazione predefiniti. È spesso utile per eseguire semplici operazioni di pulizia e pre-elaborazione su piccole quantità di dati per i quali Glue o EMR sarebbero overkilling.

Amazon EMR è la piattaforma AWS per i big data e per l'elaborazione di grandi quantità di dati utilizzando strumenti open source come Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi e Presto.

AWS Glue è un servizio di preparazione dei dati basato su Spark senza server che semplifica l'estrazione, la trasformazione e il caricamento (**ETL**) di enormi set di dati sfruttando job PySpark.

I data scientist possono anche utilizzare **AWS Glue DataBrew** per ripulire visivamente e normalizzare i dati senza scrivere codice. Inoltre, AWS Glue Studio può essere utilizzato anche per creare processi di trasformazione dei dati più articolati per Glue utilizzando una GUI user flriendly.

Analysis and Visualization

In questa fase del processo di analisi dei dati, i dati sono solitamente già puliti e aggregati in informazioni utili ai fini dell'analisi.

Una volta che tutte le informazioni richieste sono pronte è il momento di visualizzarle e analizzarle per ottenere le informazioni necessarie.

Per visualizzare i dati elaborati e il risultato dell'analisi puoi sfruttare Amazon Quicksight. **Amazon QuickSight** è un servizio di business intelligence veloce e basato su cloud che semplifica la distribuzione di informazioni dettagliate a tutti nella tua organizzazione. Essendo un servizio completamente gestito, QuickSight consente di creare e pubblicare facilmente dashboard interattivi a cui è possibile accedere da qualsiasi dispositivo e incorporati nelle applicazioni, nei portali e nei siti Web.

La visualizzazione dei dati può essere integrata con i risultati dell'apprendimento automatico al fine di fornire: analisi dei dati descrittiva, diagnostica, predittiva, prescrittiva e cognitiva.

L'analisi descrittiva risponde alla domanda: **cosa è successo?** Si concentra sul senno di poi ed è spesso chiamato data mining.

L'analisi diagnostica risponde alla domanda: **perché è successo?** Si concentra sul senno di poi e sull'intuizione. Questa forma di analisi viene utilizzata per confrontare i dati storici con altri dati provenienti da fonti diverse. Utilizzando questo metodo è possibile trovare dipendenze e schemi che possono portare alle risposte.

L'analisi predittiva risponde alla domanda: **cosa succederà?** Si concentra su intuizione e lungimiranza. Questa forma di analisi utilizza i risultati dell'analisi descrittiva e diagnostica per prevedere eventi e tendenze futuri. L'accuratezza di questo metodo dipende fortemente dalla qualità dei dati e dalla stabilità della situazione prevista.

L'analisi prescrittiva risponde alla domanda: **cosa devo fare?** Si concentra sulla previsione. Questa forma di analisi viene utilizzata per prescrivere le azioni da intraprendere sulla base dei dati forniti. Questo tipo di analisi richiede input da tutte le altre forme di analisi, combinate con regole e ottimizzazione basata su vincoli, per fare previsioni pertinenti. Il più grande vantaggio di questo modulo è che può essere automatizzato. L'apprendimento automatico lo rende possibile.

L'intelligenza cognitiva e artificiale risponde alla domanda: **quali sono le azioni consigliate?** Si concentra sulla previsione e sull'input di ipotesi. Questa forma di analisi cerca di imitare ciò che fa il cervello umano nella risoluzione dei problemi. I sistemi analitici cognitivi generano ipotesi da dati, connessioni e vincoli esistenti. Le risposte vengono fornite sotto forma di raccomandazioni e una classifica di fiducia.

Questa era la nostra rapida introduzione all'analisi dei dati su AWS. Quali progetti vorreste realizzare o avete realizzato?

Continuate a seguire **#Proud2beCloud** e rimanete sintonizzati per altri articoli su questo tema!

beSharp

Dal 2011 beSharp guida le aziende italiane sul Cloud. Dalla piccola impresa alla grande multinazionale, dal manifatturiero al terziario avanzato, aiutiamo le realtà più all'avanguardia a realizzare progetti innovativi in campo IT.

Get in touch

beSharp.it
proud2becloud@besharp.it