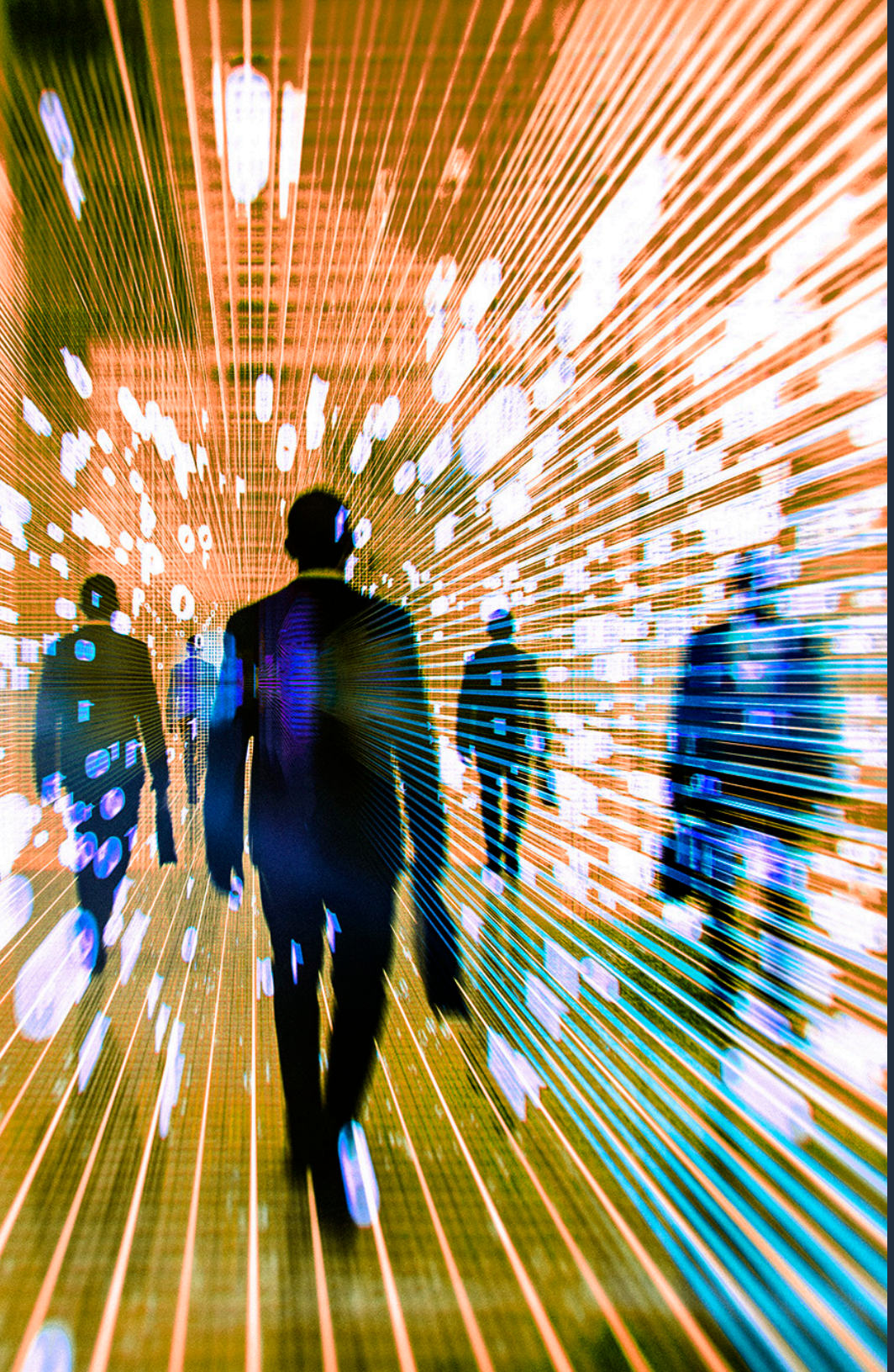
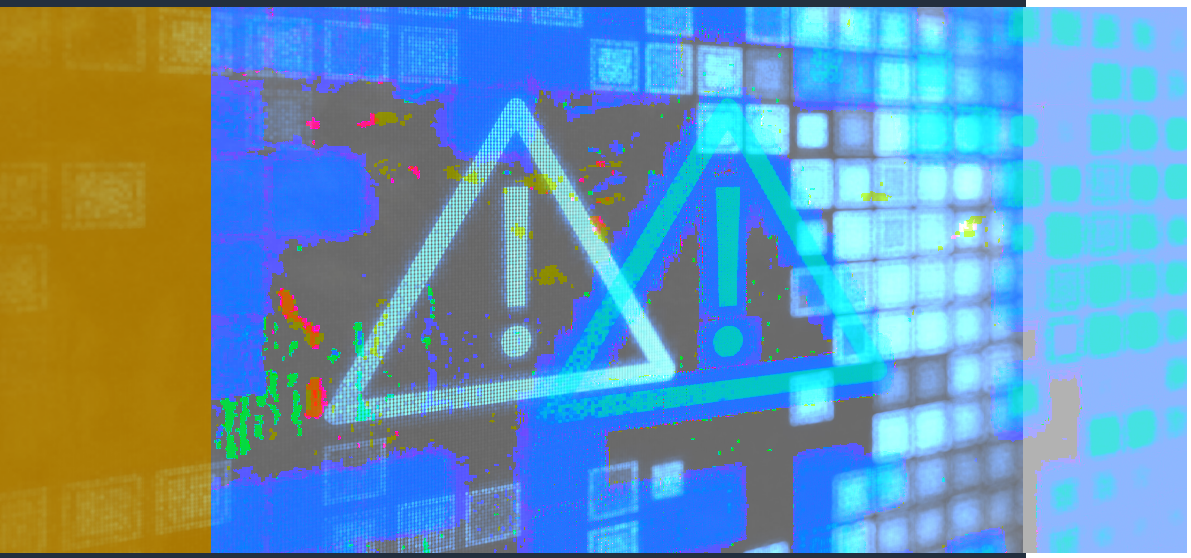


Diventare un'organizzazione basata sui dati

Joe Chung, enterprise strategist ed evangelista tecnologico presso Amazon Web Services

Ogni azienda ha un problema relativo a Big Data mascherato in modi diversi: un problema di dati



Immagina...

Hai ricevuto il report Excel settimanale via e-mail. Mentre lo controlli, ti accorgi di un'anomalia che non comprendi nei dati finanziari, nonostante la tabella pivot fornita nel rapporto ti consenta di approfondire un po' più nel dettaglio. Chiedi delucidazioni all'analista delle operazioni e l'analista risponde: "Non lo so. Fammi controllare."

Il giorno dopo, l'analista ti riferisce che l'anomalia è causata da una notevole diminuzione della produttività dell'impianto.

> "Non ha alcun senso", dici. "Puoi chiedere alle risorse umane se i giorni di malattia stanno influendo sulla produttività? O potrebbe trattarsi di un problema con l'applicazione per il monitoraggio degli orari di lavoro all'interno dello stabilimento?"
>
> "Ci vorranno settimane per raccogliere quei dati e integrarli ai dati finanziari", risponde l'analista.
>
> "Non puoi semplicemente mandarmi un dump dell'ERP e dell'applicazione degli orari, così che possa lavorarci io?"
>
> L'analista risponde: "Non ho accesso ai dati e ci vorranno settimane per inviare i ticket appositi per ottenerli."

Se questo scenario ti sembra anche troppo familiare, la tua organizzazione ha un problema con i Big Data.

La tua prima reazione potrebbe essere di considerare questa situazione come un processo di business intelligence e un limite strumentale che affligge le organizzazioni da sempre, piuttosto che un vero problema di Big Data. Senza addentrarci in un dibattito religioso sul confronto tra analisi, reporting e business intelligence, il concetto chiave è che ***ogni* organizzazione affronta problemi di Big Data**. Con l'affermarsi delle capacità dell'intelligenza artificiale e del Machine Learning, è ora più importante che mai per le aziende acquisire un maggior controllo sui propri dati e sfruttarne le potenzialità per diventare un'**organizzazione basata sui dati**.

Il mio modello mentale di riferimento per il processo di trasformazione in organizzazione basata sui dati è il sistema nervoso del corpo umano. Le estremità nervose si estendono lungo tutto il nostro corpo, inviando segnali sensoriali alla spina dorsale, che vengono poi elaborati dal nostro cervello per formulare delle reazioni. È un modello che viene emulato nelle architetture dati dotate della capacità di ricevere, elaborare e archiviare in tempo reale dati provenienti dall'interno e dall'esterno dell'azienda. Gli algoritmi di Machine Learning elaborano i segnali in tempo reale e reagiscono ad essi. Purtroppo, fin troppe organizzazioni considerano queste funzionalità utili, ma non necessarie, ritenendole applicabili esclusivamente a scenari di dati specializzati. Oppure, tentano di riproporre piattaforme di business intelligence precedenti come "data lake".

Disfunzioni dei dati

La maggior parte di noi ritiene che i problemi di Big Data riguardino il volume. La verità è che ogni organizzazione ha problemi di dati non correlati al volume, che vengono mascherati in diversi modi. Ecco alcune comuni disfunzioni dei dati che ho osservato:

Dati solitari ed eliminati

Innanzitutto, diverse organizzazioni non si rendono conto che vaste quantità di dati interessanti vengono eliminate o non sono accessibili. Alcuni esempi includono dati come l'attività dell'utente nell'applicazione (e il modo in cui potrebbe utilizzare l'applicazione in relazione ad altre applicazioni), la telemetria dell'infrastruttura che ospita l'applicazione o versioni obsolete dei dati non più compatibili con gli schemi attuali delle tabelle.

In secondo luogo, i dati vengono conservati in diverse applicazioni e data warehouse. Anche se le applicazioni non sono "grandi", quando vengono considerate nell'insieme, costituiscono una notevole entità. Perciò, quando l'azienda deve analizzare dati provenienti da molteplici fonti, si trova a dover affrontare una sfida complessa. Questo perché i dati immagazzinati in modo distribuito creano un problema di accesso. Ogni luogo in cui vengono archiviati i dati impone ruoli di accesso, regole e protocolli propri da rispettare, che rendono difficoltoso l'accesso ai dati.

"

Ogni azienda ha un problema relativo a Big Data mascherato in modi diversi."

Dati a bassa fedeltà

I sistemi aziendali tradizionali generalmente elaborano e acquisiscono solo gli stati finali, riportando tipicamente piccole istantanee temporali. Inoltre, i dati vengono elaborati in batch piuttosto che in tempo reale.
I dati possono variare sensibilmente tra un'operazione di batch e l'altra, ma i vecchi sistemi sono spesso progettati per eliminare i cambiamenti che avvengono durante le operazioni, perché non sono in grado di gestire la velocità con cui i dati cambiano.

Dati rotondi in tabelle quadrate

Molte aziende si rendono conto dell'esistenza di grandi quantità di dati che non possono essere organizzati con precisione utilizzando le tecnologie di archiviazione dei database tradizionali (per esempio immagini, dati di sensori, ecc.). In aggiunta, sono disponibili molti modi per analizzare e sfruttare le informazioni estraibili dai dati. Per esempio, lanciando una nuova iniziativa di analisi, potrebbe risultare evidente che non esiste una singola soluzione di report e visualizzazione che soddisfi tutte le necessità degli utenti. Potrebbe essere necessario considerare di fornire informazioni elaborate dagli algoritmi tramite le API, all'interno di applicazioni che utilizzano widget di visualizzazione personalizzati basati su framework JavaScript come D3.js e attraverso portali di business intelligence che sfruttano Tableau e altre soluzioni di visualizzazione.

Dati disorganizzati

I dati disorganizzati non trovano posto all'interno dei sistemi aziendali, ragion per cui esistono moduli, regole e altre metodologie di convalida, la cui finalità è quella di assicurare che i dati siano il più puliti possibile prima di venire archiviati. Ma alcuni dei dati più interessanti potrebbero non essere puliti. Quando si inizia a lavorare con dati non strutturati o dati basati su oggetti, è inevitabile trovare delle interferenze, sottoforma di grandi quantità di dati. Proprio come nel caso delle interferenze elettroniche, esistono meccanismi di filtraggio, potenziamento e amplificazione utilizzabili per ottenere i dati desiderati.
Un caso d'uso che preoccupa molto le organizzazioni riguarda l'aumento vertiginoso dei costi causati dall'invio di dati a strumenti proprietari di sicurezza, monitoraggio e aggregazione dei log. Nella maggior parte di questi casi, esiste l'opportunità di abbassare i costi filtrando ed eliminando una vasta porzione di dati di log superflui.

Se questi problemi sono noti e queste situazioni risultano familiari, è tempo che la tua organizzazione rivaluti la propria architettura e il proprio approccio analitico. Ogni azienda ha l'opportunità di adottare soluzioni analitiche appropriate (archiviazione, elaborazione, esecuzione di query, analisi, presentazioni, ecc.) per soddisfare le proprie esigenze di business e far fronte alle sfide informatiche.

Le moderne piattaforme analitiche permettono di valutare informazioni aziendali strategiche

Dopo aver deciso di affrontare il problema con i Big Data, quali sono le aspettative ragionevoli di una moderna piattaforma di analisi? Ecco quali sono le possibilità e le soluzioni disponibili al momento, da un punto di vista tecnico.

> "Le decisioni basate sui dati richiedono l'accesso ai tipi più disparati di informazioni."

Accesso a tutti i dati desiderati

Le decisioni basate sui dati richiedono l'accesso ai tipi più disparati di informazioni. Un pilota fa affidamento sulle spie e gli indicatori dell'aereo per ricevere informazioni critiche durante il volo, come altitudine, velocità di volo e consumo di carburante. Ma cosa succederebbe se questi strumenti non si trovassero tutti nello stesso posto? Magari dovrebbe andare nella cabina posteriore, chiedere informazioni via radio o, peggio ancora, richiedere l'autorizzazione per accedere ai dati. Purtroppo, questa è una realtà quotidiana nell'ambiente aziendale odierno.

Le organizzazioni più intraprendenti hanno ribaltato questo standard estraendo i dati dai sistemi in cui si trovano e memorizzandoli in un singolo posto (ovvero, il Data Lake). Anche se esistono molti esempi di società che archiviano grandi quantità di dati di un solo tipo, ne esistono sempre di più che creano Data Lake grandi come tutta l'azienda e contenenti tipi di dati provenienti da fonti differenti.

Le società di Internet come Amazon, Yahoo e Facebook hanno notato nei primi anni 2000 che le tecnologie dei database relazionali avevano raggiunto i propri limiti in termini di scalabilità e prestazioni. Amazon ha risposto con una tecnologia chiamata Dynamo, un archivio dati chiave/valore ad elevata disponibilità e scalabilità, come la tecnologia NoSQL/non relazionale. Amazon si è poi evoluta e ha sfruttato Dynamo per creare servizi come Amazon S3 e Amazon DynamoDB. Amazon S3 è ideale per le aziende che desiderano creare dei Data Lake, grazie alla sua capacità di archiviare molti tipi di dati differenti e al suo basso costo. Ci sono, ovviamente, altre soluzioni tecniche, tra cui Hadoop, ma un'importante caratteristica delle soluzioni data lake è la loro capacità di archiviare tutti i tipi di dati con una scalabilità nell'ordine di petabyte a un basso costo.

Reattività al cambiamento

I sistemi aziendali e i dati cambiano in continuazione, ma spesso i sistemi che riportano o condividono le informazioni finiscono per essere gli ultimi a cambiare. Quante volte ti è stato detto che ci sarebbero voluti 6 mesi o più per soluzioni di remediation dei dati nei data warehouse e nei report? O che i cambiamenti nei dati dei sistemi d'origine non sono ancora confluiti nei sistemi di report e che ci vorranno diversi giorni prima che quei cambiamenti si propaghino a causa del batch processing? **La velocità con cui sono disponibili i dati determina la velocità con cui si possono prendere decisioni**. Perciò, dovremmo aspettarci che i moderni sistemi di analisi siano in grado di elaborare e riportare i dati quasi in tempo reale e che siano reattivi ai cambiamenti nelle fonti di dati a monte.

> "La velocità con cui sono disponibili i dati determina la velocità con cui si possono prendere decisioni."

Il primo passo in avanti fondamentale consiste nella natura della modalità di archiviazione delle tecnologie Big Data come Amazon S3 o Hadoop. Uno dei maggiori inibitori del cambiamento in un database relazionale è rappresentato dalla modifica dello schema o dalla definizione di come i dati dovrebbero essere archiviati. Se lo schema non viene modificato, l'immissione di dati potrebbe causare l'arresto anomalo del database. Le tecnologie basate su oggetti o su file come Amazon S3 non tengono conto della struttura dei dati: essi vengono accettati nella loro forma originale, piuttosto che forzati nella struttura esistente.

L'altra sfida consiste nel fatto che un solo schema può essere attivo in un determinato momento. Anche se sono sicuro che tutti noi abbiamo visto tabelle di database denominate "2015" e "2016", questa prassi non è ideale. Le tecnologie Big Data presentano un approccio basato sulla lettura. Ciò significa che la struttura dei dati viene applicata quando vengono prelevati e non dedotta in base a come sono archiviati. Per le aziende, ciò significa che i cambiamenti dei dati dai sistemi d'origine non rappresentano un grande problema.

Il secondo fattore chiave sono le tecnologie di streaming come Amazon Kinesis e Apache Spark. La maggior parte delle aziende sposta i dati in grossi batch; tipicamente, questo accade una volta al giorno. Le tecnologie di streaming permettono di acquisire i dati in porzioni più piccole su una scala molto ampia. Per esempio, SONOS, il produttore di altoparlanti, elabora 1 miliardo di eventi ogni settimana utilizzando Amazon Kinesis. Non si dovrebbe essere costretti ad aspettare che il batch giornaliero venga completato per capire qual è la situazione dell'azienda.

Informazioni interattive dove e come le voglio

Oggi, gli utenti di un'azienda devono fare i salti mortali per comprendere le informazioni che ricevono. Magari scavando nella casella di posta per trovare un rapporto inviato come allegato. O accedendo al sistema di report per scaricare un PDF, per poi scoprire che bisogna copiare e incollare i dati in Excel per rendere le informazioni fruibili. Dobbiamo impedire che gli utenti debbano ancora superare incredibili difficoltà per ottenere i dati e le informazioni di cui necessitano. Il motto degli utenti dovrebbe essere: dati nel formato giusto, con gli strumenti giusti e al momento giusto.

Software come Tableau, Amazon QuickSight e altri hanno migliorato le cose focalizzandosi sull'esperienza utente nell'interazione con i dati. Tuttavia, ho notato che la maggior parte delle aziende richiede l'uso di molti strumenti per soddisfare le richieste degli utenti. Come ad esempio Amazon QuickSight incorporato in un portale di business intelligence contenuto in una cartella di lavoro di Tableau inviata per e-mail. AWS offre la diversità dell'archiviazione dei dati e strumenti di business intelligence attraverso un modello di pagamento a consumo. Questo permette alle organizzazioni di sperimentare con molti strumenti di business intelligence differenti senza dover fare grandi investimenti in infrastrutture e licenze.

Dobbiamo impedire che gli utenti debbano ancora superare incredibili difficoltà per ottenere i dati e le informazioni di cui necessitano."

Una figura da tenere bene in considerazione è quella del data scientist della tua azienda. I notebook Jupyter sono ormai ampiamente utilizzati nella comunità della Data Science e sono costituiti in parte da content management, in parte da esecuzione di codice e in parte da visualizzazione. Si tratta di uno strumento molto potente per condividere conoscenze, documentare ed eseguire algoritmi di Machine Learning. Amazon SageMaker è un ambiente notebook gestito che si occupa del lavoro pesante al posto tuo e del tuo data scientist.

L'intelligenza artificiale come parte integrante dell'azienda

Oggi, l'intelligenza artificiale e il Machine Learning sono considerati tecnologie d'avanguardia. E per un buon motivo. I progressi nei framework di Machine Learning uniti all'uso di server specializzati che sfruttano unità di elaborazione grafica (GPU) permettono nuove funzionalità di ogni tipo, come la guida autonoma. Ovviamente, per l'addestramento dei modelli di Machine Learning sono necessarie grandi quantità di dati (da cui i Data Lake di cui ho discusso in precedenza). Le organizzazioni stanno già iniziando a sfruttare queste funzionalità di IA/ML per ottenere risultati finora impossibili, come la possibilità di diagnosi sanitarie più accurate sulla base di retinografie o di prevedere interruzioni dei servizi e guasti hardware. Le organizzazioni possono ottenere il massimo dai propri sistemi di IA/ML, lasciando a AWS il grosso del lavoro; non si tratta di fantascienza ma di una realtà operativa esistente già oggi.

Un ultimo punto che vorrei sottolineare è che anche il miglior algoritmo al mondo è inutile a meno che non possa essere integrato con i processi aziendali. Spesso, analizzare correttamente i dati o creare modelli di Data Science è la parte più semplice del lavoro. La difficoltà sta nell'integrare questi modelli nel motore per le polizze di assicurazione o nella piattaforma di vendita al dettaglio, dal momento che questi sistemi tipicamente non dispongono della capacità di integrare fonti di dati o API esterne. Questa è una fantastica opportunità per considerare la migrazione di questi sistemi al cloud, in modo da approfittare di tutti i servizi disponibili per modernizzarli e ristrutturarli.

L'organizzazione per una corretta analisi

Sviluppare funzionalità analitiche avanzate per la tua azienda non è solo una questione di tecnologia. Spesso, la più grande sfida per le organizzazioni è l'organizzazione stessa, con i suoi processi, la sua governance e i suoi dipendenti. Perciò, quali sono i fattori fondamentali per il successo quando si investe nell'analisi?

"Un ultimo punto che vorrei sottolineare è che anche il miglior algoritmo del mondo è inutile a meno che non possa essere integrato con i processi aziendali."

Inizia da un Centro di eccellenza per l'analisi

Per passare dalla strategia e dall'intento al progresso significativo, uno dei primi passi consiste nell'identificazione e nella scelta di un leader e di un team che promuovano il cambiamento, oltre che alla creazione di un Centro di Eccellenza (COE) per l'analisi. Di solito si parte da un team ridotto, con qualche ruolo multifunzionale per le prime fasi, per poi crescere man mano che aumentano le necessità da soddisfare.

Molte grandi aziende vantano già organizzazioni consolidate per i servizi condivisi, che si occupano del reporting o della business intelligence. Queste organizzazioni popolano il COE per l'analisi di ruoli tecnici e amministrativi. Analogamente alle organizzazioni IT, il loro scopo non è solo quello di fornire personale specializzato, ma anche quello di agire come leader e sponsor nell'impegno verso l'obiettivo comune. Col tempo, infatti, queste organizzazioni per i servizi condivisi di reporting dovranno evolversi per adattarsi o divenire parte del COE per l'analisi. I ruoli di partenza sono spesso quelli di data engineer, data architect, business intelligence analyst e data scientist. Il gruppo deve essere guidato da qualcuno capace di lavorare in più organizzazioni, unità operative e gruppi amministrativi, come il settore finanziario e IT.

Soddisfare tutte le necessità dei clienti

Uno dei primi passi nel cambiamento di mentalità che le organizzazioni devono compiere riguarda il passaggio dall'approccio "devi usare la nostra soluzione di reporting senza lamentarti" a "quali sono le tue necessità di analisi e come possiamo aiutarti a soddisfarle?" Le organizzazioni che si occupano del reporting di servizi condivisi sono spesso semplici diffusori di report e non si trovano nella posizione di poter rispondere alle difficili problematiche sollevate da dipendenti, business lead e clienti.

Perciò, nel creare un nuovo COE analitico, è importante stabilire dei principi per il gruppo, in base ai quali definire le sue modalità nell'agire e prendere decisioni.

Il COE analitico dovrà servire due tipi di clienti:

- **I consumer di dati e analisi**: i responsabili delle decisioni, i Data Scientist, gli analisti di business intelligence (BI) e gli sviluppatori. Per questi clienti solitamente è importante poter accedere velocemente a informazioni strategiche e dati, oltre a poter contare su strumenti e servizi di qualità per elaborare e presentare i dati.
- **I produttori di dati**: i proprietari di applicazioni, infrastrutture e dispositivi che invieranno dati alla piattaforma. Questi clienti richiedono servizi come la capacità di pubblicare facilmente i propri dati sulla piattaforma di analisi e la definizione di un contratto dati. Ciò include il modello di dominio dei dati, la frequenza degli aggiornamenti e la definizione di criteri come per esempio un criterio di sicurezza che stabilisca chi può accedere ai dati.

> "Non si tratta solo di avere a disposizione gli strumenti più avanzati, ma di aiutare i clienti (produttori o utenti) a ottenere ciò di cui hanno bisogno con facilità."

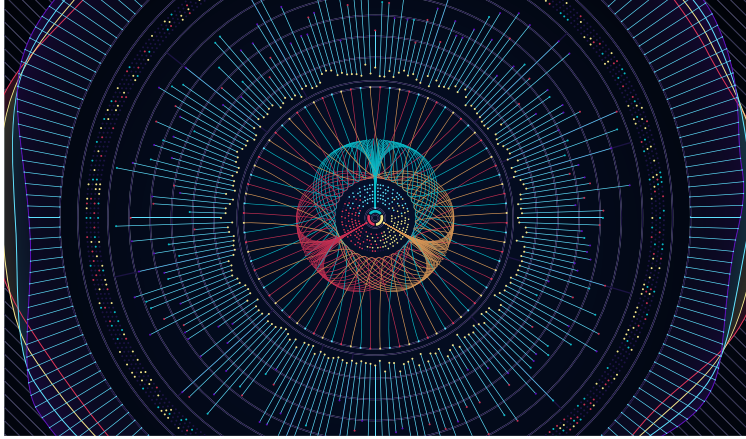
Informazioni sull'autore

Joe Chung Enterprise Strategist ed evangelista tecnologico presso Amazon Web Services.

Le funzioni di analisi e la piattaforma devono servire entrambi i tipi di cliente: se le loro necessità non vengono soddisfatte, allora lo sforzo analitico non porterà valore all'impresa. Per questo motivo, è cruciale disporre di un meccanismo per comprendere le necessità di questi due tipi di clienti su un insieme potenzialmente molto vasto e diversificato di unità operative e utenti tipo. Diverse organizzazioni creano comitati consultivi o collaborano con alcuni stakeholder chiave per promuovere tali esigenze. Non esiste una sola risposta corretta, ma avere a disposizione un meccanismo che consideri la voce del cliente e dia priorità alle sue esigenze è fondamentale.

Ripensare il COE

Un COE per l'analisi fornisce un insieme specializzato di servizi cloud focalizzati sulla soddisfazione delle esigenze di analisi. In passato, le organizzazioni di reporting e BI offrivano spesso un'unica soluzione per soddisfare le necessità di tutti (una strategia a taglia unica). In quest'era di tecnologie di Big Data in rapida evoluzione, visualizzazioni avanzate, decisioni automatizzate, intelligenza artificiale e Machine Learning è semplicemente impossibile affidarsi a un solo stack tecnologico. Non si tratta solo di avere a disposizione gli strumenti più avanzati, ma di aiutare i clienti (produttori o utenti) a ottenere ciò di cui hanno bisogno con facilità.



I COE corrono il rischio di diventare dei servizi concierge. Sebbene ciò potrebbe funzionare per certi tipi di richieste, il COE può rapidamente venire sovraccaricato e riempito di richieste arretrate se non si dispone di meccanismi self-service scalabili o di processi trasparenti di governance e di assegnazione della priorità. I COE per l'analisi devono progettare e architettare una piattaforma dati self-service sicura, funzionale e scalabile con un ecosistema di tecnologie in costante evoluzione per elaborare, analizzare e presentare le informazioni strategiche.

Anche se la trasformazione in un'organizzazione basata sui dati non avverrà rapidamente, definire con precisione gli obiettivi, organizzare un piano su come soddisfare i clienti e fornire ai team il potere di offrire il valore giusto al momento giusto sono tutti passi nella giusta direzione.