

Le principali cause di fallimento di un progetto Data-Driven e come il Cloud le mitiga

6 Novembre 2024 - 6 min. read

AI

Cloud adoption

Data and Analytics

Data Ingestion

Data Security and Governance

Generative AI

Machine Learning

Introduzione

Nel mondo dell'innovazione, il concetto di *fail-fast* si riferisce alla capacità di identificare e risolvere rapidamente i problemi, riducendo al minimo gli investimenti su approcci che potrebbero non dare i risultati sperati. Questa mentalità, sempre più essenziale nei progetti tecnologici, permette di risparmiare tempo e risorse, consentendo di testare ipotesi e strategie in modo agile. Nell'ambito dei progetti data-driven, dove il valore può essere estratto solo attraverso iterazioni e continue ottimizzazioni, la filosofia del *fail-fast* rappresenta un metodo per sperimentare, apprendere e migliorare senza penalizzare l'intero progetto.

Il progetto data-driven non va visto solo nella sua componente di data analysis o di elaborazione del dato tramite algoritmi di Machine Learning o l'implementazione della Generative AI. Le fasi di un progetto data-driven partono dalla pianificazione e gestione del flusso di dati dall'origine (come fonti di dati grezzi) fino all'elaborazione e analisi finali. In generale, possiamo raggruppare le fasi di questa tipologia di progetti in tre macro blocchi: data ingestion, data management e data value. In ogni punto si possono riscontrare insidie che portano progetti interi a naufragare. A supporto di una strategia ben definita, il Cloud ci aiuta a massimizzare il valore dei dati e a ridurre i rischi di errori o inefficienze che possono portare a fallimenti.

Il Cloud, dunque, è l'abilitatore perfetto di questo approccio, grazie alla sua flessibilità, all'elevata scalabilità e alla presenza di servizi innovativi. A differenza delle infrastrutture on-

premises, il cloud permette di avviare Proof of Concept (PoC) a costi iniziali ridotti e senza l'onere di investimenti infrastrutturali significativi. Questo consente alle aziende di testare in modo rapido e a basso costo, riducendo l'impatto economico di eventuali errori nelle varie fasi dei progetti data-centrici.

Come il cloud migliora la qualità e la disponibilità dei dati

La qualità e disponibilità dei dati rappresentano elementi essenziali per il successo dei progetti data-driven. Un progetto che parte con dati di scarsa qualità o con accesso limitato rischia infatti di fallire non portando a risultati significativi, con impatti economici negativi derivanti dai costi di sviluppo. Nella fase di *Data Ingestion*, il cloud, grazie ai servizi erogati e alla possibilità di mettere in gioco automatismi out-of-the-box, consente di migliorare la qualità e la disponibilità dei dati e, al tempo stesso, di ridurre i costi associati, semplificando il lavoro dei data engineer.

Servizi come AWS Glue consentono di configurare pipeline di dati automatizzate, con funzionalità avanzate di data-cleansing (o data cleaning) e data-transformation, anche in tempo reale.

Tramite tecnologie come Amazon S3/Glacier, il cloud rende possibile l'implementazione di data lake e data warehouse durevoli e scalabili che permettono di centralizzare grandi volumi di dati grezzi provenienti da più fonti a basso costo, garantendo facilità di accesso e gestione. I data engineer possono configurare facilmente pipeline di ingestion che acquisiscono dati in tempo reale o batch, assicurando una disponibilità continua e accessibile a tutti i team che ne necessitano.

In un ambiente on-premises, garantire un accesso costante e sicuro ai dati richiederebbe un notevole sforzo infrastrutturale e, in caso di errori di processo, progettuali o di concetto, il costo del fallimento aumenterebbe significativamente.

Data la natura a consumo del Cloud si elimina il rischio di sovradimensionare le risorse, un problema comune nei sistemi on-premises, dove il fallimento comporta un notevole spreco di risorse infrastrutturali già acquistate. Di conseguenza, diventa efficiente testare rapidamente i processi di ingestion su ambienti temporanei. Questo significa che è possibile iterare facilmente sulle pipeline, correggere errori senza impatti permanenti e contenere i costi associati a configurazioni errate.

Uno dei pilastri fondamentali dei progetti data-driven: la governance del dato

La governance del dato rappresenta uno dei capisaldi dei progetti data-driven e racchiude più aspetti che toccano sicurezza, accesso e compliance normativa. La governance riguarda la gestione e il controllo dell'intero ciclo di vita dei dati, compresa l'autorizzazione all'accesso, la tracciabilità e il monitoraggio delle attività. Il cloud computing gioca un ruolo centrale nell'ottimizzare la governance dei dati, facilitando la gestione dei costi associati a possibili fallimenti nella fase di *Data Management*.

Nel contesto on-premises, la governance dei dati è spesso complessa e costosa da gestire: richiede infrastrutture dedicate, team di sicurezza dedicati e configurazioni manuali che aumentano i tempi e i costi. Con il cloud, invece, le aziende possono accedere a strumenti di governance integrati che semplificano la definizione delle policy, la gestione dei permessi e il monitoraggio delle attività, riducendo significativamente il rischio di errori. Piattaforme come AWS offrono servizi di *Identity and Access Management* (IAM), crittografia avanzata (KMS) e log di accesso (CloudTrail), che centralizzano la gestione delle autorizzazioni rendendola più semplice.

Una delle forze principali degli ambienti Cloud, che abbiamo già toccato, è l'automazione dei processi, in questo caso di governance. Ad esempio, con i servizi di audit logging (come AWS CloudTrail), i dati sensibili sono continuamente monitorati, e ogni accesso viene registrato in modo che si possa intervenire tempestivamente su eventuali anomalie o violazioni di policy. Questa tracciabilità automatizzata facilita la compliance e minimizza l'impatto economico di eventuali errori o configurazioni sbagliate. Con il cloud, anche eventuali problemi di accesso possono essere risolti rapidamente senza compromettere la sicurezza dei dati o richiedere interventi manuali.

Il Cloud a supporto delle decisioni strategiche basate sul valore dai dati

Il concetto di *Data Value* si riferisce al processo attraverso il quale i dati grezzi vengono trasformati in insight significativi e attuabili, essenziali per guidare decisioni strategiche. Tuttavia, le scelte analitiche sbagliate o progettazioni inefficaci possono generare costi elevati, prolungare i tempi di sviluppo e, nei casi peggiori, portare al fallimento dell'intero progetto.

Nella fase di progettazione di un flusso analitico, il data scientist o l'analista deve prendere decisioni cruciali su come trattare i dati: quali modelli di machine learning utilizzare, quali variabili includere e come gestire i processi di training e validazione dei modelli. Con il cloud, è possibile testare diverse configurazioni di modelli analitici in parallelo, sfruttando l'elaborazione scalabile e pagando solo per il tempo di utilizzo. Ad esempio, strumenti come Amazon SageMaker, permettono di sperimentare in modo rapido e iterativo, abbassando il costo di errori o ipotesi sbagliate e ottimizzando i modelli in modo continuo.

Il cloud facilita la gestione del flusso dei dati attraverso strumenti di orchestrazione avanzata, come AWS Step Functions, che permettono di progettare pipeline di elaborazione dati modulari e flessibili. Nel contesto on-premises, eventuali errori nel flow richiederebbero spesso la riprogettazione dell'intero processo, con conseguenti perdite di tempo e risorse.

Questa modularità permette di adattare rapidamente il flusso dei dati alle necessità del business, garantendo flessibilità anche in caso di cambiamenti strategici.

Conclusione

Il cloud è un abilitatore per progetti che richiederebbero infrastrutture troppo complesse o costose se sviluppate on-premises. Consentendo accesso a risorse on-demand e a consumo, il cloud apre la strada a iniziative come il disaster recovery a basso costo, l'archiviazione massiva, l'intelligenza artificiale, la genomica e la simulazione ambientale, che necessitano di un'elaborazione intensiva e/o scalabile.

In particolare nei progetti data-driven, il cloud facilita la gestione dei dati con strumenti integrati che supportano data ingestion, governance e qualità del dato.

L'approccio agile, basato sulla sperimentazione veloce e sul concetto di fail-fast, riduce le barriere economiche e di complessità, rendendo possibili progetti ambiziosi per ogni tipologia di azienda, indipendentemente dalla sua dimensione.

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!



Nicola Ferrari

Cloud Infrastructure Line Manager @ beSharp e AWS authorized instructor champion. Vivo la vita “un livello alla volta”. Ottengo i miei superpoteri raccogliendo caffeina nascosta qua e là nella mia mappa quotidiana. Sono un Internet surfer professionale (e ho visto tutto l’Internet per intero... almeno due volte!) e un appassionato di tecnologia, computer e networking. Costruire grandi cose IT - tutte precise e ordinate - contribuisce alla mia missione principale: la ricerca della perfezione!

Copyright © 2011-2024 by beSharp spa - P.IVA IT02415160189