AUTOPSIA DI UN'ANOMALIA INVISIBILE

AXIANTE

UNA STORIA DI FANTASIA, MODELLATA SUL QUOTIDIANO: QUEL MESSAGGIO SU TEAMS CHE INTERROMPE IL PRIMO CAFFÈ DEL MATTINO, UN NUMERO INSPIEGABILE CHE COMPARE NEL REPORT... E LA SENSAZIONE CHE QUALCOSA STIA TRABALLANDO, ANCHE SE TECNICAMENTE NON C'È NULLA DI ROTTO.

Autopsia di un’anomalia invisibile

Ogni settimana guardo i numeri scorrere silenziosi nelle dashboard direzionali. Per altri possono sembrare solo valori, per me sono molto di più: sono il linguaggio con cui l’azienda misura il proprio passo, prende decisioni, valuta strategie. Sono un data analyst e dietro a quei numeri c’è un lavoro estremamente complesso: pipeline, controlli, orchestrazioni, accordi tra sistemi e persone. Un ecosistema progettato per garantire affidabilità e trasparenza.

Chi lavora nel mondo dei dati lo sa: basta un piccolo errore o una mancata comunicazione per far crollare in un attimo la fiducia nel reporting aziendale.

Eppure so bene che anche gli ecosistemi più solidi vivono di equilibri delicati. Basta un dettaglio imprevisto, una comunicazione mancata, e l’ingranaggio perfetto mostra subito una crepa.

È da una crepa così che è iniziata questa storia.

LA DOCCIA FREDDA

A metà mattina ricevo un messaggio improvviso su Teams. È Clara, controller di direzione. Nessun ticket ufficiale, nessuna mail con copie al management: solo uno screenshot della dashboard di sintesi.

“Ciao Lorenzo, qui i numeri non tornano. Vendite settimanali Core A troppo basse.”

> *Per altri possono sembrare solo valori, per me sono molto di più: sono il linguaggio con cui l’azienda misura il proprio passo, prende decisioni, valuta strategie*

Quella dashboard non è un semplice grafico: è lo strumento direzionale che alimenta il segment reporting per il board. Basta poco per incrinarne l’affidabilità.

Apro subito la vista segnalata e lo vedo anch’io: i valori sono più bassi del previsto. Non è un problema di filtro o di selezione, il dato è proprio quello. Nessun alert dai sistemi, nessuna pipeline fallita, nessun log sospetto.

Mi passo una mano tra i capelli, guardando il cursore lampeggiare sullo schermo. Il sotto-conteggio è del 12% sulle vendite settimanali Core A: abbastanza da falsare previsioni, bonus e decisioni strategiche.

Appoggio il caffè sulla scrivania e, quasi senza accorgermene, ho già sei tab di log aperti. È il classico momento da incubo in cui pensi: se tutti gli alert sono verdi, vuol dire che qualcosa di grosso ti sta sfuggendo.

Capisco immediatamente che mi serve un confronto più ampio. Chiamo Michela, la Data Engineer. Nessun preambolo: «Ciao Michela, c'è un'anomalia sulle vendite settimanali, core business. Ho verificato i log, nulla di irregolare. Mi aiuti a rivedere insieme i possibili punti deboli?».

Lei sospira appena, ma accetta senza esitazioni. Decidiamo di affrontare l'analisi in modo sistematico, con metodo: verifica su Airflow, Spark, dbt, orchestrazione e quality check. Nessun salto nel buio, ogni ipotesi va testata e chiusa.

I PRIMI SOSPETTI

Condivido lo schermo e le mostro la dashboard che Clara mi ha segnalato.

«Se i numeri sono più bassi,» dico, «la prima cosa che mi viene in mente è un problema a monte: dati mancanti, incompletezza del caricamento, qualcosa di simile.»

Michela annuisce, ma controlla subito i log. «Ho già verificato gli arrivi giornalieri. Tutti i file sono presenti, nessun drop. I controlli di qualità a monte hanno validato i dataset: valori non nulli, campi obbligatori compilati, formati corretti. Se fosse saltato qualcosa lì, l'avremmo intercettato.»

«Ok,» rispondo, «quindi qualità a monte esclusa. Passiamo al tema latenze: dati tardivi o arrivati fuori sequenza. Se avessimo ricevuto flussi incompleti, magari con eventi mancanti, avremmo un sotto-conto delle vendite.»

Lei apre il monitor dei watermark.
«I controlli di watermark sono allineati. Le pipeline accettano fino a 48 ore di ritardo e non ci sono backlog. Gli arrivi sono stati puntuali e ordinati.»

Prendo nota. Mi mordo l'interno della guancia mentre scorro la lista: niente, tutto in ordine.
«E se fosse un problema di eventi invertiti? Clickstream, log, cose arrivate fuori ordine e quindi conteggiate male?» tento.

«Abbiamo implementato la gestione del reorder sugli eventi time-based. Anche qui nessun alert: i processi hanno ricostruito correttamente l'ordine temporale.»

Resto un attimo in silenzio, poi rilancio: «Allora vediamo la deduplica. Se qualche transazione fosse stata filtrata come duplicata per errore, il volume sarebbe più basso».

«Controllato,» ribatte. «Le logiche di idempotenza funzionano: stessa chiave, stessa transazione. Non ci sono stati scarti anomali, i conteggi sono nella norma.»
Appoggio la penna sulla scrivania. Sospiro, ma non sono scoraggiato. «Perfetto. Quindi: qualità dei dati in ingresso, latenze, reorder, deduplica... tutte ipotesi scartate.»

«Esatto,» conferma Michela. «Il problema non è in questi strati. Dobbiamo proseguire, passare a dimensioni e calendari.»
La lista delle possibili cause si sta già assottigliando. **E con essa cresce quella tensione sorda che conosciamo bene, quando il puzzle continua a non incastrarsi.**

DIMENSIONI E CALENDARI SOTTO ESAME

«So che non possiamo saltare passaggi,» dico. «Un'altra possibile fonte di errore è nelle dimensioni. Se qualche attributo di prodotto o cliente fosse stato gestito male in una Slowly Changing Dimension, potremmo avere valori spostati o mal attribuiti.»

Michela apre i log di aggiornamento delle dimensioni. «Le SCD sono presidiate con chiavi surrogate e storicizzazione. Ho controllato: nessun update fuori sequenza, nessuna nuova versione aperta senza chiusura della precedente. I prodotti della Categoria Core A sono associati correttamente, senza buchi di validità.»

Annuisco. «Quindi possiamo escludere problemi di attribuzione. Rimane la questione dei calendari. Se ci fosse stato uno shift di timezone o un'incoerenza sui periodi di aggregazione, i numeri settimanali potrebbero risultare falsati.»

Lei consulta il job che consolida i calendari di riferimento. «Ho verificato: gli eventi sono normalizzati tutti su UTC e ricalibrati sulla timezone aziendale. La settimana di riferimento è quella corretta, lunedì-domenica. Non ci sono state modifiche ai calendari fiscali né variazioni di cut-off.»
Insisto su un dettaglio: «E le festività locali? Qualche variazione nel calendario operativo delle Legal Entity?».

«Controllato,» mi dice. «Il calendario operativo delle Entità è allineato al repository master. Nessuna incongruenza. Le aggregazioni sono coerenti.»
Il nodo non si scioglie, ma ogni ipotesi esclusa rende più chiaro il quadro. Michela mi guarda di sfuggita, quasi come a dire: "qui non c'è nulla, dobbiamo scavare altrove".

«Bene,» concludo. «Allora anche dimensioni e calendari li mettiamo da parte. Andiamo avanti: il prossimo passo è rivedere definizioni KPI, orchestrazione e processi schedulati.

Magari il problema non è nei dati, ma nel modo in cui li sintetizziamo o li calcoliamo.»

Il percorso prosegue, strato dopo strato.

KPI, ORCHESTRAZIONE E LIMITI INFRASTRUTTURALI

«Se i dati di base sono a posto,» rifletto, «il problema potrebbe essere nelle definizioni dei KPI. Magari un cambiamento di logica di calcolo che non abbiamo intercettato.»

Michela si muove veloce tra repository e documentazione. «Le definizioni dei KPI sono versionate. Non ci sono state modifiche nelle ultime settimane e la logica di fatturato netto è identica a quella già validata. Nessuna regressione né override manuale.»

«Allora guardiamo alle dipendenze,» continuo. «Se una pipeline upstream avesse ritardato, il nostro processo potrebbe aver calcolato su dati incompleti.»
Lei apre la console dell'orchestratore e io trattengo il respiro.

«Tutti i DAG hanno girato regolarmente. Non ci sono task falliti né retry. Le dipendenze critiche erano soddisfatte al momento dell'esecuzione. Nessun warning.» spiega.

Prendo appunti, e sono sempre più teso. «Ok. E se invece fosse un tema infrastrutturale? Quote superate, job interrotti per limiti di risorse, costi troppo alti che hanno bloccato qualcosa?»

Michela scuote la testa e mi mostra i grafici di monitoraggio. «Le risorse sono nei limiti, nessun job è stato terminato per OOM o per quota. I costi di esecuzione sono stabili, nessun alert da parte dell'infrastruttura cloud.»

Un'altra ipotesi sfuma.
«Quindi: KPI stabili, orchestrazione ok, infrastruttura regolare. Anche questo blocco è escluso.»

«Sì,» conferma lei. «Restano pochi strati da indagare: metadati e sicurezza. Poi, se anche lì non troviamo nulla, dobbiamo guardare a monte, alle sorgenti.»

Accidenti! So che quel "guardare a monte" è l'ipotesi più delicata. Significa mettere in discussione l'unico pezzo che non possiamo blindare con codice.

METADATI E SICUREZZA SOTTO LA LENTE

Apro il catalogo dati, quasi anticipando la prossima domanda. «Potrebbe essere un problema di metadati. Se il catalogo non fosse aggiornato, potremmo leggere tabelle o campi non più validi. In quel caso, i numeri sarebbero distorti.»

Michela segue il mio flusso e controlla le ultime esecuzioni di aggiornamento del metastore. «Il catalogo è allineato: aggiornato in automatico a ogni deploy delle pipeline. Non risultano campi deprecati né lineage interrotte. Le viste della dashboard puntano tutte a oggetti coerenti con l'ultimo schema.»

«Quindi niente metadati obsoleti. E se invece fosse un problema di permessi? Magari la Legal Entity Core A ha restrizioni e stiamo leggendo dati parziali per motivi di sicurezza.»

La collega apre la sezione di access management. «I permessi non sono cambiati. La dashboard direzionale legge con privilegi di servizio centralizzati, non soggetti a restrizioni per singola entità. Nessuna policy è stata aggiornata negli ultimi rilasci. Anche qui, nessuna anomalia.»

Mi appoggio allo schienale e incrocio le braccia. «Allora rimane l'ultima possibilità: le sorgenti. Se non è qualità, né latenze, né dimensioni, né KPI, né metadati... significa che qualcosa a monte è cambiato senza che ci venisse comunicato.»

Michela annuisce, consapevole che stiamo toccando il punto più critico. Poi, aggiunge: «Sì. Dobbiamo verificare i contratti con le sorgenti. È l'unico ambito che non possiamo controllare in autonomia. Se qualcuno avesse introdotto un nuovo tipo di documento senza avvisarci, i nostri processi non lo riconoscerebbero. E il fatturato risulterebbe certamente più basso».
Il cerchio si chiude. Ogni ipotesi tecnica è stata esclusa. Rimane la più *umana*: la comunicazione tardiva.

Dopo giorni di verifiche, controlli e confronti, io e Michela arriviamo finalmente al punto cruciale: le sorgenti. Tutto il resto l'abbiamo escluso con metodo.
Michela apre il flusso di caricamento dalla Legal Entity segnalata e fa girare un controllo sulle tipologie di documento presenti. Dopo pochi minuti, appare il risultato sullo schermo: un codice nuovo, mai visto prima.

«Eccolo,» dice. «Un nuovo Tipo Documento di Vendita. Non era previsto, non era nel contratto, e non ci è stato comunicato.»

Scorro l'elenco e vedo subito il dettaglio. «E riguarda solo questa Legal Entity... e solo alcuni prodotti di una segmentazione particolare. Esattamente quelli che Clara ha notato sul report.»

Il puzzle si ricompone. Non c'erano bug nelle pipeline, né errori nei calcoli. Era stata l'introduzione tardiva di un nuovo documento operativo, non allineato al modello dati, a produrre un sotto-conteggio del fatturato.
Restiamo in silenzio per qualche istante. Poi Michela commenta con pragmatismo: «Questa è la parte che nessun sistema può coprire: la mancanza di comunicazione. Tutto il resto, lo avevamo previsto e blindato. Ma se a monte cambiano regole senza avvisarci, la discrepanza è inevitabile».

Annuisco. «È la dimostrazione che i sistemi sono solidi, ma non onniscienti. Possono garantire qualità, consistenza, coerenza... ma non possono prevedere il futuro.»

OLTRE LA TECNICA

Io e Michela abbiamo percorso ogni strato, scartando ipotesi con metodo e verificando la solidità di ciò che abbiamo costruito. La causa del problema si è rivelata esterna al sistema: un cambiamento non comunicato. Non un difetto tecnico, ma un limite umano.

Il valore di questo percorso non è stato soltanto trovare la radice dell'anomalia, ma mostrare quanto ogni livello sia pensato per reggere l'imprevisto. E ricordare che i sistemi migliori non sostituiscono mai il dialogo tra le persone ma, al contrario, lo richiedono.

Forse è proprio questo il mestiere di chi lavora sui dati: non essere solo manutentori di pipeline, ma guardiani della coerenza, traduttori del trait d'union tra numeri e decisioni. Ogni giorno costruiamo processi robusti, sapendo che il punto più fragile rimane quello che non possiamo automatizzare: la comunicazione. **Perché l'affidabilità dei dati non nasce solo dal codice, ma anche dalla disciplina della comunicazione.** E il futuro, per quanto si cerchi di anticiparlo, rimane sempre la parte più difficile da prevedere.

www.axiante.com