



Analisi post-imprevisto

Interruzione di aprile 2022

Questa traduzione è fornita solo per comodità dell'utente. In caso di ambiguità o conflitto tra le traduzioni, prevale la versione originale in inglese.

Lettera dei nostri co-fondatori e co-CEO

Desideriamo comunicare di essere al corrente dell'interruzione che ha interrotto il servizio per i clienti all'inizio del mese. Comprendiamo che i nostri prodotti sono fondamentali per la vostra azienda e non ci assumiamo questa responsabilità alla leggera. Ci consideriamo i primi responsabili. Responsabili completi. Ai clienti interessati, vogliamo dire che stiamo lavorando per riconquistare la loro fiducia.

i In Atlassian, uno dei nostri valori fondamentali è «Promuovi un'azienda aperta, non solo a parole». Incarniamo questo valore anche discutendo apertamente degli imprevisti e usandoli come opportunità di apprendimento. Pubblichiamo questa revisione post-imprevisto per i nostri clienti, la nostra Atlassian Community e la community tecnica più in generale. Atlassian è orgogliosa del suo [processo di gestione degli imprevisti](#) che pone l'accento sul fatto che una cultura irreprensibile e l'individuazione di modi per migliorare i sistemi e processi tecnici sono fondamentali per fornire servizi affidabili e di alta qualità. Pur facendo del nostro meglio per evitare qualsiasi tipo di imprevisto, siamo anche dell'idea che gli imprevisti rappresentino un modo efficace per migliorare.

La piattaforma cloud di Atlassian ci consente di soddisfare le diverse esigenze dei nostri oltre 200.000 clienti cloud di ogni dimensione e settore. Prima di questo imprevisto, il tempo di attività del nostro cloud si è costantemente tenuto al 99,9% superando i tempi di attività previsti dagli SLA. Abbiamo effettuato investimenti a lungo termine nella nostra piattaforma e in una serie di funzionalità della piattaforma centralizzata, con un'infrastruttura scalabile e una cadenza costante di miglioramenti della sicurezza.

Rivolgiamo un ringraziamento ai nostri clienti e ai nostri partner per la costante fiducia e collaborazione. Ci auguriamo che le informazioni e le azioni delineate in questo documento dimostrino che Atlassian continuerà a fornire una piattaforma cloud del massimo livello e a costituire un valido portafoglio di prodotti per soddisfare le esigenze di ogni team.



-Scott e Mike

Riepilogo esecutivo

Martedì 5 aprile 2022, a partire dalle 7:38 UTC, 775 clienti Atlassian hanno perso la possibilità di accedere ai loro prodotti Atlassian. Per un sottoinsieme di questi clienti, l'interruzione è durata addirittura 14 giorni. Il ripristino di un primo gruppo di clienti è avvenuto l'8 aprile e tutti i siti dei clienti sono stati progressivamente ripristinati entro il 18 aprile.

L'imprevisto non è stato il risultato di un attacco informatico e non vi è stato alcun accesso non autorizzato ai dati dei clienti. Atlassian ha un programma completo di [gestione dei dati](#), con tanto di SLA pubblicati, e ha in passato costantemente superato tali SLA.

Per quanto l'imprevisto sia stato grave, nessun cliente ha perso più di cinque minuti di dati. Inoltre, oltre il 99,6% dei nostri clienti e utenti ha continuato a utilizzare i nostri prodotti cloud senza interruzioni durante le attività di ripristino.



In questo documento, i clienti i cui siti sono stati eliminati nell'ambito di questo imprevisto sono indicati come clienti «interessati». Questo PIR fornisce i dettagli esatti dell'imprevisto, delinea le azioni che abbiamo intrapreso per il ripristino e descrive come eviteremo il ripetersi di situazioni come questa in futuro. Forniamo un riepilogo generale dell'imprevisto in questa sezione, con ulteriori dettagli nel resto del documento.

Cosa è successo?

Nel 2021 abbiamo completato l'acquisizione e l'integrazione di un'app Atlassian standalone per Jira Service Management e Jira Software denominata "Insight – Asset Management". La funzionalità di questa app standalone era quindi nativa all'interno di Jira Service Management e non era più disponibile per Jira Software. Per questo motivo, avevamo bisogno di eliminare l'app legacy standalone sui siti dei clienti su cui era installata. I nostri team di ingegneri hanno utilizzato uno script e un processo esistenti per eliminare istanze di questa applicazione standalone, ma si sono verificati due problemi:

- **Mancanza di comunicazione.** C'è stata una mancanza di comunicazione tra il team che ha richiesto l'eliminazione e il team che l'ha eseguita. Anziché fornire gli ID dell'app da contrassegnare per la disattivazione, il team ha fornito gli ID dell'intero sito cloud in cui le app dovevano essere disattivate.

- **Avvisi di sistema insufficienti.** L'API utilizzata per eseguire l'eliminazione ha accettato sia gli identificatori del sito che dell'app e ha dato per scontato che l'input fosse corretto: ciò significava che se l>ID di un sito veniva trasmesso, un sito sarebbe stato eliminato; se fosse stato trasmesso l>ID di un'app, un'app sarebbe stata eliminata. Non è stato visualizzato alcun segnale di avviso per confermare il tipo di eliminazione (sito o app) richiesta.

Lo script che è stato eseguito era stato sottoposto al nostro processo di revisione paritaria standard, basato sull'endpoint chiamato e sulla modalità della chiamata. Non effettuava controlli incrociati sugli ID del sito cloud forniti, per verificare se si riferivano all'Insight App o all'intero sito. Il problema è stato che lo script conteneva l>ID dell'intero sito di un cliente e ha causato la cancellazione immediata di 883 siti (riconducibili a 775 clienti) tra le 07:38 UTC e le 08:01 UTC di martedì 5 aprile 2022. *Vedi «Cosa è successo»*

Come abbiamo risposto?

Una volta che il 5 aprile alle 08:17 UTC l'imprevisto è stato confermato, abbiamo attivato il nostro processo di gestione degli imprevisti gravi e abbiamo formato un team interfunzionale di gestione degli imprevisti. Il team globale di risposta agli imprevisti ha lavorato 24 ore su 24, 7 giorni su 7, per tutta la durata dell'imprevisto, fino a quando tutti i siti non sono stati ripristinati. Inoltre, i responsabili della gestione degli imprevisti si sono incontrati ogni tre ore per coordinare i flussi di lavoro.

All'inizio, ci siamo resi conto delle numerose serie che comportava il ripristino simultaneo di centinaia di clienti con più prodotti.

All'inizio dell'imprevisto, sapevamo esattamente quali siti erano stati interessati e la nostra priorità è stata di metterci in comunicazione con il responsabile approvato per ogni sito interessato per informarlo dell'interruzione.

Tuttavia, alcuni recapiti dei clienti erano stati eliminati. Ciò significava che i clienti non potevano presentare richieste di assistenza come avrebbero fatto normalmente. Ciò significava anche che non avevamo accesso immediato ai contatti chiave dei clienti.

Per ulteriori dettagli, vedere «Panoramica generale dei flussi di lavoro di ripristino»

Cosa stiamo facendo per prevenire situazioni come questa in futuro?

Abbiamo intrapreso una serie di azioni immediate e ci impegniamo ad apportare modifiche per evitare il ripetersi di questa situazione in futuro. Ecco quattro aree specifiche in cui abbiamo apportato o apporteremo modifiche significative:

1. **Istituzione di «eliminazioni temporanee» universali su tutti i sistemi.** In generale, un'eliminazione di questo tipo dovrebbe essere vietata o avere più livelli di protezione per evitare errori, come l'implementazione graduale e un piano di rollback testato per le «eliminazioni temporanee». Impediremo a livello globale l'eliminazione dei dati e metadati dei clienti che non sono stati sottoposti a un processo di eliminazione temporanea.
2. **Accelerazione del nostro programma di ripristino di emergenza (Disaster Recovery, DR) per automatizzare il ripristino in caso di eliminazione multi-sito e multi-prodotto per un gruppo più ampio di clienti.** Sfrutteremo l'automazione e gli insegnamenti tratti da questo imprevisto per accelerare il programma di DR e raggiungere l'obiettivo RTO (Recovery Time Objective) definito nella nostra politica per questa portata di incidenti. Eseguiremo regolarmente procedure di ripristino di emergenza che prevedono il ripristino di tutti i prodotti per un'ampia serie di siti.
3. **Revisione del processo di gestione degli imprevisti per imprevisti su larga scala.** Miglioreremo la nostra procedura operativa standard per imprevisti su larga scala e la metteremo in pratica con simulazioni. Aggiungeremo la nostra formazione e gli strumenti per gestire il gran numero di team che lavorano in parallelo.
4. **Creazione di una guida operativa per le comunicazioni su larga scala.** Riconosceremo gli imprevisti per tempo, attraverso più canali. Rilascieremo comunicazioni pubbliche sugli incidenti nel giro di poche ore. Per raggiungere meglio i clienti interessati, miglioreremo il backup dei contatti chiave e gli strumenti di supporto retrofit per consentire ai clienti senza un URL o un ID Atlassian valido di entrare in contatto diretto con il nostro team di supporto tecnico.

L'elenco completo delle azioni è fornito in dettaglio nella revisione post-impvosto completa riportata di seguito. *Vedi «Come miglioreremo»*

Sommario

Panoramica sull'architettura cloud di Atlassian	Pagina 7
<ul style="list-style-type: none">• Architettura di cloud hosting di Atlassian• Architettura dei servizi distribuiti• Architettura multi-tenant• Provisioning e ciclo di vita del tenant• Programma di ripristino di emergenza<ul style="list-style-type: none">○ Resilienza○ Capacità di ripristino dell'archiviazione di servizio○ Capacità di ripristino automatizzata per più siti e più prodotti	
Cosa è successo, timeline e ripristino	Pagina 13
<ul style="list-style-type: none">• Cosa è successo• Come ci siamo coordinati• Timeline dell'imprevisto• Panoramica generale dei flussi di lavoro di ripristino<ul style="list-style-type: none">○ Flusso di lavoro 1: rilevamento, avvio del ripristino e identificazione del nostro approccio○ Flusso di lavoro 2: ripristino anticipato e approccio Restoration 1○ Flusso di lavoro 3: ripristino accelerato e approccio Restoration 2○ Perdita di dati minima in seguito al ripristino dei siti eliminati	
Comunicazioni sugli imprevisti	Pagina 21
<ul style="list-style-type: none">• Cosa è successo	
Esperienza di assistenza e comunicazione con i clienti	Pagina 23
<ul style="list-style-type: none">• In che modo ne ha risentito il supporto per i nostri clienti?• Come abbiamo risposto?	
Come miglioreremo?	Pagina 25
<ul style="list-style-type: none">• Insegnamento 1: le «eliminazioni temporanee» devono essere universali su tutti i sistemi• Insegnamento 2: nell'ambito del programma DR, automatizzare il ripristino per eventi di eliminazione multi-sito e multi-prodotto per un gruppo più ampio di clienti• Insegnamento 3: migliorare il processo di gestione degli imprevisti per eventi su larga scala• Insegnamento 4: migliorare i nostri processi di comunicazione	
Osservazioni conclusive	Pagina 31

Panoramica sull'architettura cloud di Atlassian

Per comprendere i fattori che hanno contribuito all'imprevisto descritto in questo documento, è utile comprendere innanzitutto l'architettura di distribuzione dei prodotti, dei servizi e dell'infrastruttura di Atlassian.

Architettura di cloud hosting di Atlassian

Atlassian utilizza Amazon Web Services (AWS) come fornitore di servizi cloud e le sue strutture di data center ad alta disponibilità in [più regioni del mondo](#). Ogni regione AWS è una posizione geografica separata con più gruppi di data center, isolati e fisicamente separati, noti come zone di disponibilità (AZ).

Sfruttiamo i servizi di calcolo, archiviazione, rete e dati di AWS per sviluppare i nostri prodotti e i componenti della piattaforma, che ci consentono di utilizzare le funzionalità di ridondanza offerte da AWS, come zone di disponibilità e regioni.

Architettura dei servizi distribuiti

Con questa architettura AWS, gestiamo in hosting una serie di servizi di piattaforma e prodotto utilizzati nelle nostre soluzioni. Ciò include funzionalità di piattaforma condivise e utilizzate in più prodotti Atlassian, come Media, Identity, Commerce, esperienze come il nostro Editor, nonché funzionalità di prodotto specifiche, come il servizio Jira Issue e Confluence Analytics.

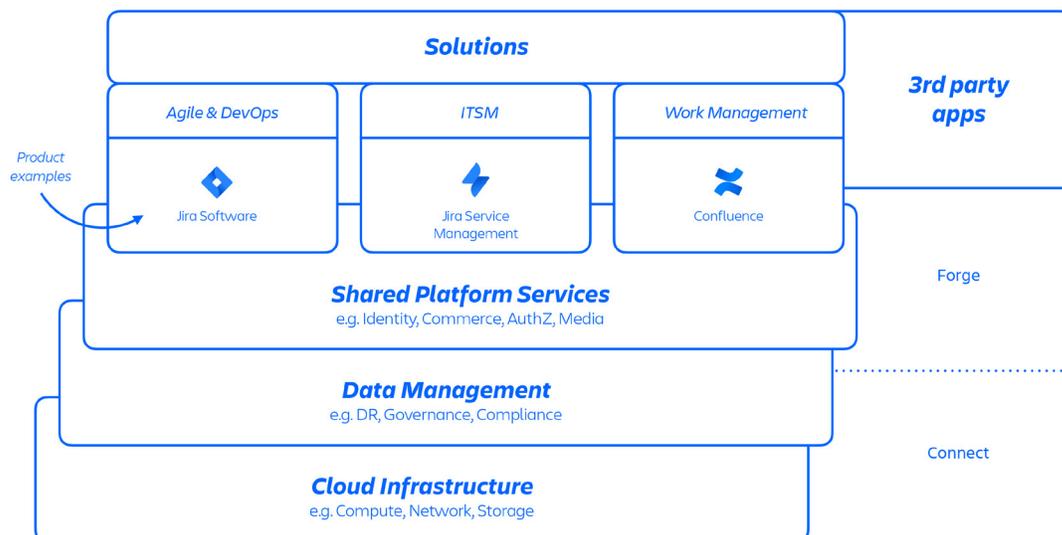


Figura 1: architettura della piattaforma Atlassian.

Gli sviluppatori Atlassian assicurano il provisioning di questi servizi tramite una piattaforma come servizio (PaaS) sviluppata internamente, denominata Micros, che orchestra automaticamente la distribuzione di servizi condivisi, infrastruttura, archivi dati e le relative funzionalità di gestione, tra cui i requisiti di sicurezza e controllo della conformità (vedere la *Figura 1* qui sopra). In genere, un prodotto Atlassian è costituito da più servizi «containerizzati» distribuiti su AWS utilizzando Micros. I prodotti Atlassian utilizzano funzionalità di base della piattaforma (vedere la *Figura 2* di seguito) come il routing delle richieste, gli archivi di oggetti binari, l'autenticazione/autorizzazione, i contenuti generati dagli utenti (UGC) transazionali e gli archivi di relazioni tra entità, data lake, comune registrazione, tracciamento delle richieste, osservabilità e analisi. Questi microservizi sono costruiti utilizzando stack tecnici approvati e standardizzati a livello di piattaforma:

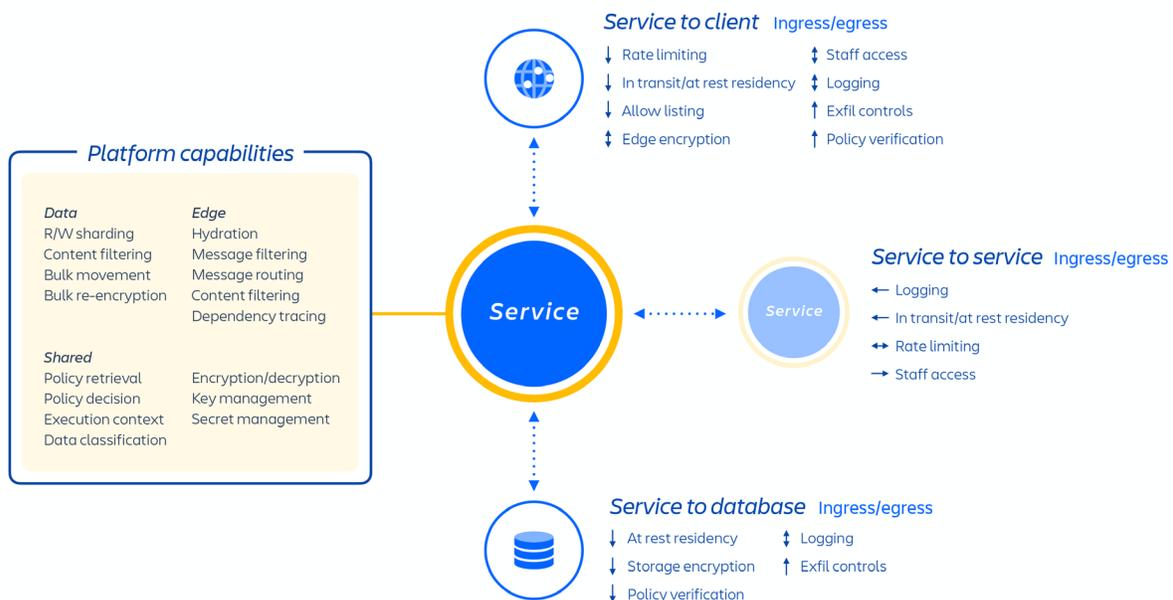


Figura 2: panoramica dei microservizi Atlassian.

Architettura multi-tenant

Oltre alla nostra infrastruttura cloud, abbiamo sviluppato e gestiamo un'architettura di microservizi multi-tenant accanto a una piattaforma condivisa che supporta i nostri prodotti. In un'architettura multi-tenant, un unico servizio serve più clienti, inclusi database e istanze di calcolo necessari per eseguire i prodotti cloud. Ogni frammento (essenzialmente un contenitore - vedere la *Figura 3* di seguito) contiene i dati per più tenant, ma i dati di ciascun tenant sono isolati e inaccessibili agli altri tenant.

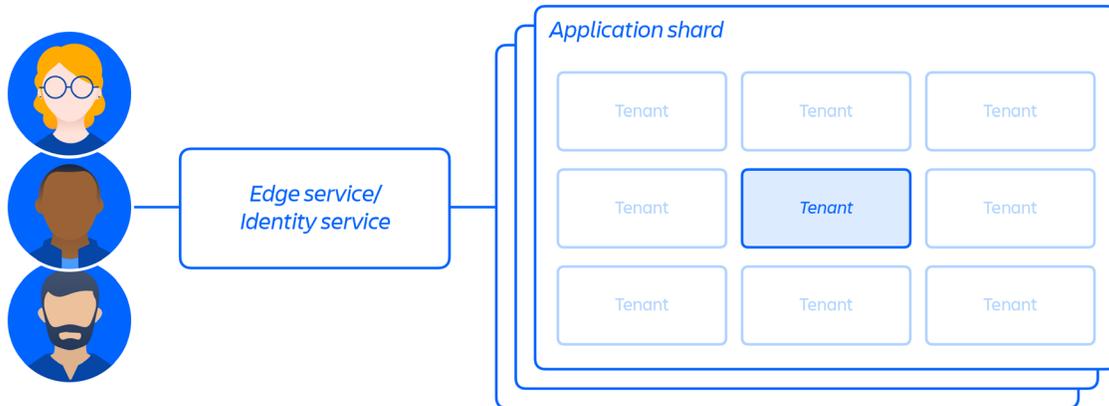


Figura 3: come archiviamo i dati in un'architettura multi-tenant.

Provisioning e ciclo di vita del tenant

Quando viene eseguito il provisioning di un nuovo cliente, una serie di eventi attivano l'orchestrazione dei servizi distribuiti e il provisioning di archivi dati. Questi eventi possono essere generalmente mappati a uno dei sette passaggi del ciclo di vita:

- 1 I sistemi di e-commerce vengono immediatamente aggiornati con i metadati più recenti e le informazioni di controllo degli accessi per quel cliente, quindi un sistema di orchestrazione del provisioning allinea lo «stato delle risorse assegnate» allo stato della licenza attraverso una serie di eventi tenant e di prodotto.

Eventi tenant

Questi eventi riguardano il tenant nel suo complesso e possono essere i seguenti:

- Creazione: un tenant viene creato e utilizzato per siti completamente nuovi
- Distruzione: un intero tenant viene eliminato

Eventi sui prodotti

- Attivazione: dopo l'attivazione di prodotti concessi in licenza o app di terze parti
- Disattivazione: dopo la disattivazione di determinati prodotti o app
- Sospensione: la sospensione di un determinato prodotto esistente che disattiva l'accesso a un determinato sito di proprietà
- Annullamento della sospensione: annullamento della sospensione di un determinato prodotto esistente che consente l'accesso a un sito di proprietà

Aggiornamento della licenza: contiene informazioni sul numero di postazioni di licenza per un determinato prodotto e sul suo stato (attivo/inattivo)

- 2 Creazione del sito del cliente e attivazione del set corretto di prodotti per il cliente. Il concetto di sito è il container di più prodotti concessi in licenza a un determinato cliente. (ad esempio Confluence e Jira Software per `<nome del sito>.atlassian.net`). Questo (vedere la *Figura 4* di seguito) è un punto importante da comprendere nel contesto di questo report, poiché è proprio il container del sito che è stato eliminato in questo imprevisto e il concetto di sito è discusso in tutto il documento.

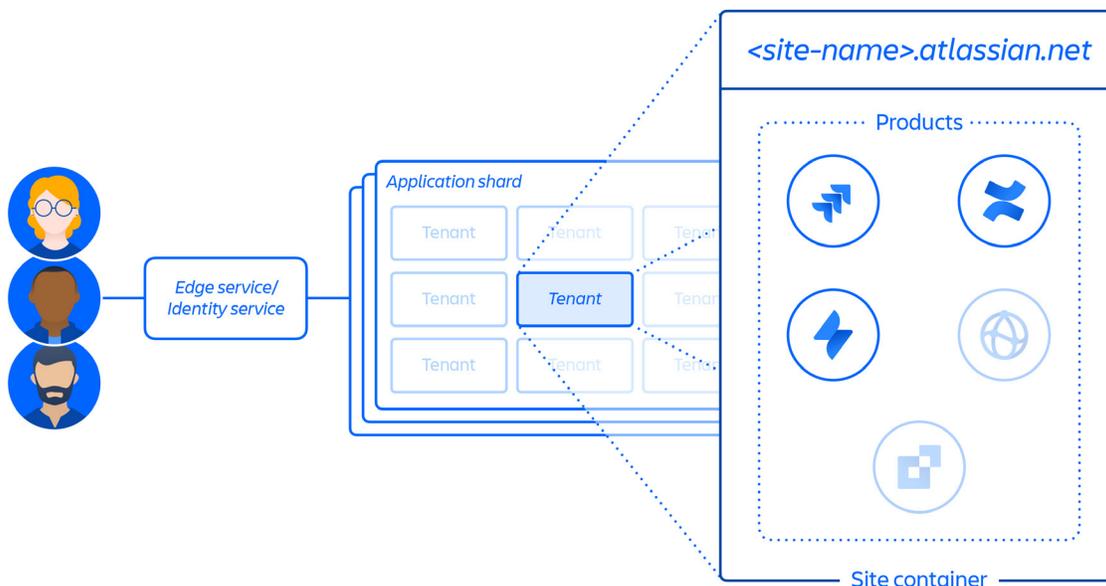


Figura 4: Panoramica del container del sito.

- 3 Provisioning di prodotti all'interno del sito del cliente nella regione designata.

Quando viene eseguito il provisioning di un prodotto, la maggior parte dei relativi contenuti sarà conservata in hosting vicino al punto in cui gli utenti vi accedono. Per ottimizzare le prestazioni del prodotto, non limitiamo lo spostamento dei dati quando sono conservati in hosting a livello globale e potremmo spostare i dati da una regione all'altra se necessario.

Per alcuni prodotti, offriamo anche la residenza dei dati. La residenza dei dati consente ai clienti di scegliere se i dati di prodotto sono distribuiti a livello globale o conservati in una delle nostre posizioni geografiche definite.

- 4 Creazione e archiviazione della configurazione e dei metadati principali del sito e del/i prodotto/i del cliente.

- 5 Creazione e archiviazione dei dati di identità del sito e dei prodotti, come gli utenti, i gruppi, le autorizzazioni, ecc.
- 6 Provisioning dei database di prodotti all'interno di un sito, ad es. famiglia di prodotti Jira, Confluence, Compass, Atlas.
- 7 Provisioning delle app con licenza del/i prodotto/i.

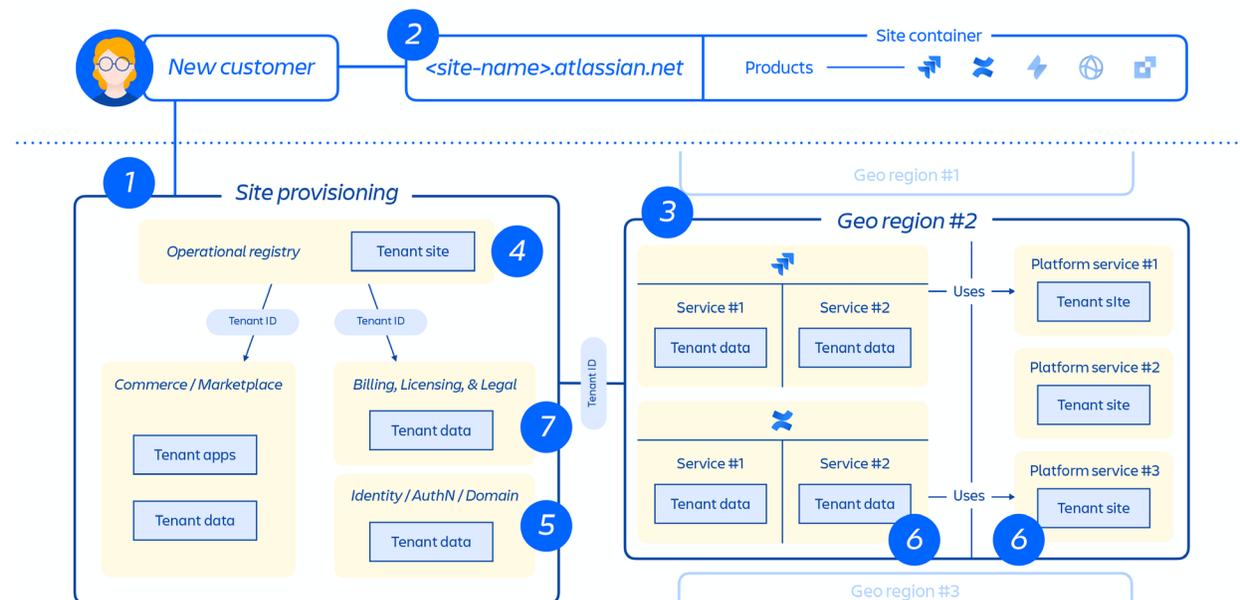


Figura 5: Panoramica del modo in cui viene eseguito il provisioning del sito del cliente nella nostra architettura distribuita.

La Figura 5 qui sopra mostra come il sito di un cliente viene distribuito nella nostra architettura distribuita, non solo in un singolo database o archivio. Ciò include più posizioni fisiche e logiche in cui sono archiviati metadati, dati di configurazione, dati di prodotto, dati della piattaforma e altre informazioni sul sito correlate.

Programma di ripristino di emergenza

Il nostro programma di [ripristino di emergenza](#) (DR) comprende tutte le azioni da noi intraprese per assicurare la resilienza in caso di errori dell'infrastruttura e la capacità di ripristino dell'archiviazione dei servizi a partire dai backup. Due concetti importanti per comprendere i programmi di ripristino di emergenza sono:

- **Obiettivo tempo di ripristino (Recovery Time Objective, RTO):** quanto velocemente è possibile ripristinare e restituire i dati a un cliente in caso di emergenza?
- **Obiettivo punto di ripristino (Recovery Point Objective, RPO):** quanto sono recenti i dati recuperati dopo il ripristino da un backup? Quanti dati andranno persi dall'ultimo backup?

Durante questo imprevisto, non abbiamo raggiunto il nostro RTO, ma abbiamo realizzato il nostro RPO.

Resilienza

Ci prepariamo per i guasti a livello di infrastruttura, ad esempio la perdita di un intero database, servizio o zona di disponibilità AWS. Questa preparazione include la replica di dati e servizi su più zone di disponibilità e test di failover regolari.

Capacità di ripristino dell'archiviazione di servizio

Ci prepariamo inoltre per il ripristino in caso di danneggiamento dei dati di archiviazione del servizio a causa di rischi come ransomware, malintenzionati, difetti del software ed errori operativi. Questa preparazione include backup immutabili e test di ripristino del backup dell'archiviazione del servizio. Siamo in grado di prendere qualsiasi archivio dati individuale e ripristinarlo a un momento precedente.

Capacità di ripristino automatizzata per più siti e più prodotti

Al momento dell'imprevisto, non avevamo la possibilità di selezionare un ampio set di siti di clienti e, a partire dai backup, ripristinare tutti i loro prodotti interconnessi a un momento precedente.

Le nostre capacità si sono concentrate su infrastruttura, danneggiamento dei dati, eventi riguardanti singoli servizi o eliminazione di singoli siti. In passato, abbiamo dovuto affrontare e testare questo tipo di guasti. La cancellazione a livello di sito non aveva runbook che potessero essere rapidamente automatizzati per la portata di questo evento che richiedeva il ricorso coordinato a strumenti e automazione in tutti i prodotti e servizi.

Nelle sezioni seguenti approfondiremo questi aspetti complessi e le azioni che stiamo intraprendendo in Atlassian per evolvere e ottimizzare la nostra capacità di mantenere questa architettura su larga scala.

Cosa è successo, timeline e ripristino

Cosa è successo

Nel 2021 abbiamo completato l'integrazione di un'app Atlassian standalone per Jira Service Management e Jira Software, denominata "Insight – Asset Management". La funzionalità di questa app standalone era quindi nativa all'interno di Jira Service Management e non era più disponibile per Jira Software. Per questo motivo, avevamo bisogno di eliminare l'app legacy standalone sui siti dei clienti su cui era installata. I nostri team di ingegneri hanno utilizzato uno script e un processo esistenti per eliminare istanze di questa applicazione standalone.

Tuttavia, sono emersi due problemi critici:

- **Mancanza di comunicazione.** C'è stata una mancanza di comunicazione tra il team che ha richiesto l'eliminazione e il team che l'ha eseguita. Anziché fornire gli ID dell'app da contrassegnare per la disattivazione, il team ha fornito gli ID dell'intero sito cloud in cui le app dovevano essere disattivate.
- **Avvisi di sistema insufficienti.** L'API utilizzata per eseguire l'eliminazione accetta sia gli identificatori del sito che dell'app e dà per scontato che l'input fosse corretto: ciò significa che se l'ID di un sito viene trasmesso, un sito viene eliminato; se viene trasmesso l'ID di un'app, un'app viene eliminata. Non è stato visualizzato alcun segnale di avviso per confermare il tipo di eliminazione (sito o app) richiesta.

Lo script che è stato eseguito era stato sottoposto al nostro processo di revisione paritaria standard, basato sull'endpoint chiamato e sulla modalità della chiamata. Non effettuava controlli incrociati sugli ID del sito cloud forniti, per verificare se si riferivano all'app o all'intero sito. Lo script è stato testato a livello di staging in base ai nostri processi di gestione delle modifiche standard, tuttavia, non avrebbe rilevato che gli ID immessi non erano corretti in quanto gli ID non esistevano nell'ambiente di staging.

Quando è stato eseguito in produzione, lo script inizialmente è stato eseguito su 30 siti. La prima esecuzione in produzione è riuscita e ha eliminato l'app Insight da quei 30 siti senza effetti collaterali. Tuttavia, gli ID per quei 30 siti erano stati acquisiti prima dell'evento di comunicazione errata e includevano gli ID corretti dell'app Insight.

Lo script per la successiva esecuzione in produzione includeva gli ID del sito al posto degli ID dell'app Insight ed è stato eseguito su un set di 883 siti. L'esecuzione dello script è iniziata il 5 aprile alle 07:38 UTC ed è stata completata alle 08:01 UTC. Lo script ha eliminato i siti in sequenza in base all'elenco di input, quindi il sito del primo cliente è stato eliminato poco dopo l'inizio dell'esecuzione dello script alle 07:38 UTC. Il risultato

è stata una cancellazione immediata degli 883 siti, senza alcun segnale di avvertimento per i nostri team di ingegneri.

I seguenti prodotti Atlassian sono risultati non disponibili per i clienti interessati: famiglia di prodotti Jira, Confluence, Atlassian Access, Opsgenie e Statuspage.

Non appena abbiamo saputo dell'imprevisto, i nostri team si sono concentrati sul ripristino per tutti i clienti interessati. Sul momento, abbiamo stimato che il numero di siti interessati fosse circa 700 (883 siti totali sono stati interessati, ma abbiamo sottratto i siti di proprietà di Atlassian). Dei 700, una parte significativa era costituita da account inattivi, gratuiti o di piccole dimensioni con un numero ridotto di utenti attivi. Sulla base di ciò, inizialmente abbiamo stimato il numero approssimativo di clienti interessati in circa 400.

Ora abbiamo una visione molto più accurata e, per una trasparenza completa in base alla definizione ufficiale di cliente Atlassian, 775 clienti sono stati colpiti dall'interruzione. Tuttavia, la maggior parte degli utenti era inclusa nella stima iniziale di 400 clienti. L'interruzione è durata fino a 14 giorni per un sottogruppo di questi clienti, con il primo gruppo di clienti ripristinato l'8 aprile e tutti i clienti ripristinati dopo il 18 aprile.

Come ci siamo coordinati

La prima richiesta di assistenza è stata inviata da un cliente interessato alle 07:46 UTC del 5 aprile. Il nostro monitoraggio interno non ha rilevato la presenza di un problema perché i siti erano stati eliminati attraverso un flusso di lavoro standard. Alle 08:17 UTC, abbiamo attivato il nostro processo di gestione degli imprevisti gravi, formando un team interfunzionale di gestione degli imprevisti e in sette minuti, alle 08:24 UTC, la situazione è stata riclassificata Critica. Alle 08:53 UTC, il nostro team ha confermato che la richiesta di assistenza e l'esecuzione dello script erano correlati. Alle 12:38 UTC, avendo compreso la complessità del ripristino, abbiamo assegnato il massimo livello di gravità all'imprevisto.

Il team di gestione degli imprevisti era composto da persone provenienti da più team di Atlassian, tra cui ingegneria, assistenza clienti, gestione dei programmi, comunicazioni e molti altri ancora. Il team principale si è riunito ogni tre ore per la durata dell'imprevisto fino a quando tutti i siti non sono stati ripristinati, convalidati e restituiti ai clienti. Per gestire l'avanzamento del ripristino, abbiamo creato un nuovo progetto Jira, SITE, e un flusso di lavoro per tenere traccia dei ripristini sito per sito tra più team (progettazione, gestione dei programmi, supporto, ecc.). Questo approccio ha consentito a tutti i team di individuare e monitorare facilmente i problemi relativi a qualsiasi ripristino del singolo sito.

L'8 aprile alle 03:30 UTC, e per tutta la durata dell'imprevisto, abbiamo anche implementato un blocco del codice in tutta la funzione di ingegneria. Questo ci ha permesso di concentrarci sul ripristino dei clienti, eliminare il rischio che modifiche causassero incongruenze nei dati dei clienti, ridurre al minimo il rischio di altre interruzioni e ridurre la probabilità che modifiche non correlate distraessero il team dal ripristino.

Timeline dell'imprevisto

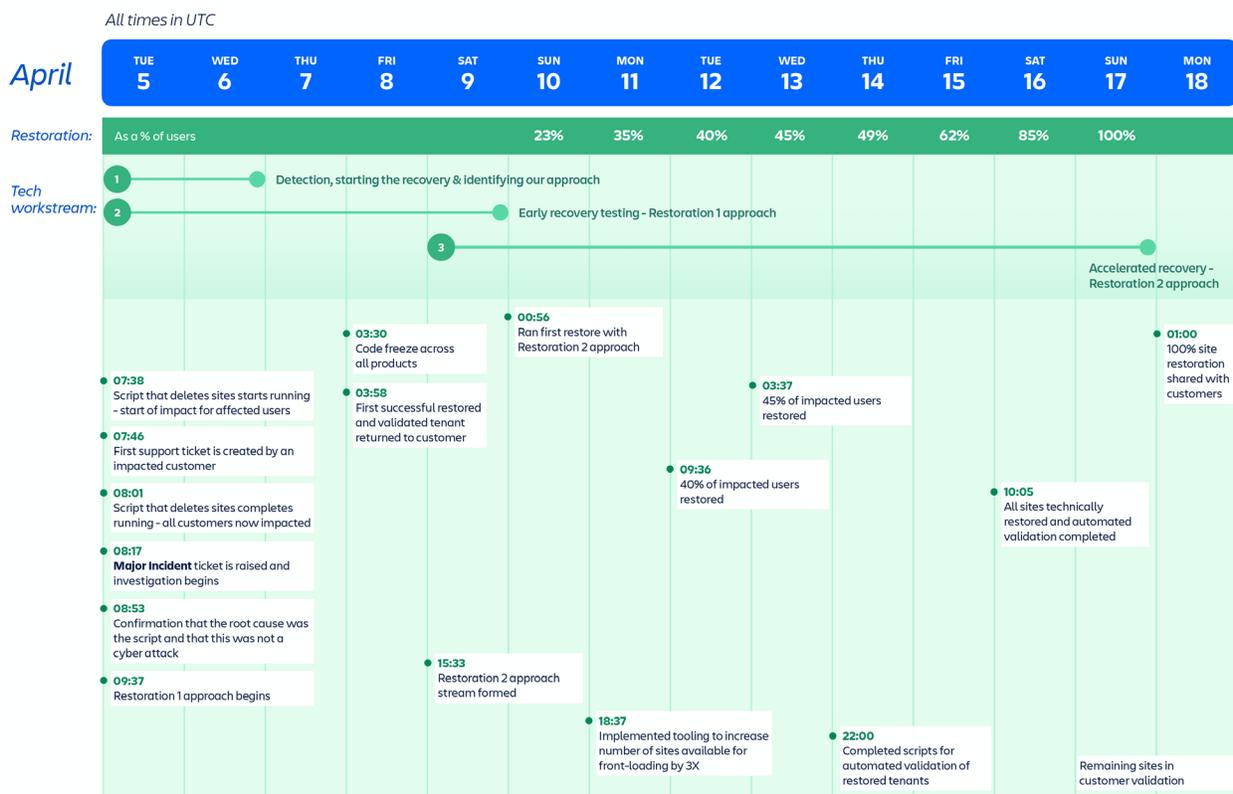


Figura 6: Timeline dell'imprevisto e milestone chiave del ripristino.

Panoramica generale dei flussi di lavoro di ripristino

Il ripristino è stato eseguito come tre flussi di lavoro principali: rilevamento, ripristino anticipato e accelerazione. Sebbene ciascun flusso di lavoro venga descritto separatamente qui di seguito, durante il ripristino il lavoro è stato svolto in parallelo su tutti i flussi di lavoro.

Flusso di lavoro 1: rilevamento, avvio del ripristino e identificazione del nostro approccio

Timestamp: giorni 1-2 (5-6 aprile)

Alle 08:53 UTC del 5 aprile, abbiamo riscontrato che lo script dell'app Insight ha causato l'eliminazione dei siti. Abbiamo verificato che non si è trattato di un atto dannoso interno né di un attacco informatico. I team competenti per l'infrastruttura di prodotti e piattaforme sono stati chiamati e coinvolti nell'imprevisto.

All'inizio dell'imprevisto, abbiamo preso atto dei seguenti fatti:

- Il ripristino di centinaia di siti eliminati è un processo complesso e in più fasi (descritto nella precedente sezione sull'architettura) che richiede molti team e più giorni per essere completato correttamente.
- Avevamo la possibilità di ripristinare un singolo sito, ma non avevamo creato la capacità e i processi per ripristinarne un gran numero contemporaneamente.

Di conseguenza, abbiamo dovuto parallelizzare e automatizzare in misura considerevole il processo di ripristino per aiutare i clienti interessati a riottenere l'accesso ai prodotti Atlassian il più rapidamente possibile.

Il Flusso di lavoro 1 ha coinvolto un gran numero di team di sviluppo impegnati nelle seguenti attività:

- Individuazione ed esecuzione di passaggi di ripristino per lotti di siti nella pipeline.
- Scrittura e miglioramento dell'automazione per consentire al/i team di eseguire i passaggi di ripristino per un maggior numero di siti in un batch.

Flusso di lavoro 2: ripristino anticipato e approccio Restoration 1

Timestamp: giorni 1-4 (5-9 aprile)

Abbiamo capito che cosa aveva causato la cancellazione del sito il 5 aprile alle 08:53 UTC, cioè entro un'ora dall'esecuzione dello script. Abbiamo inoltre individuato il processo di ripristino che era stato precedentemente utilizzato per recuperare un piccolo numero

di siti in produzione. Tuttavia, il processo di ripristino per i siti eliminati su tale scala non era definito con precisione.

Per muoverci rapidamente, abbiamo diviso le prime fasi dell'imprevisto su due gruppi di lavoro:

- Un gruppo di lavoro manuale, che ha convalidato i passaggi richiesti ed eseguito manualmente il processo di ripristino per un numero limitato di siti.
- Un gruppo di lavoro automatizzato, che ha adottato il processo di ripristino esistente e ha sviluppato l'automazione per eseguire in sicurezza i passaggi su lotti di siti di maggiori dimensioni.

Panoramica dell'approccio Restoration 1 (vedere la *Figura 7* di seguito):

- È stata necessaria la creazione di un nuovo sito per ogni sito eliminato, seguito da ogni prodotto, servizio e archivio dati a valle che necessitava di un ripristino dei dati.
- Il nuovo sito sarebbe stato fornito con nuovi identificatori come `cloudId`. Questi identificatori sono tutti considerati immutabili, il che significa che molti sistemi li integrano nei record di dati. Di conseguenza, se questi identificatori cambiavano, dovevamo aggiornare grandi quantità di dati, il che è particolarmente problematico per le app ecosistemiche di terze parti.
- La modifica di un nuovo sito per replicare lo stato del sito eliminato presentava dipendenze complesse e spesso imprevedibili tra un passaggio e l'altro.

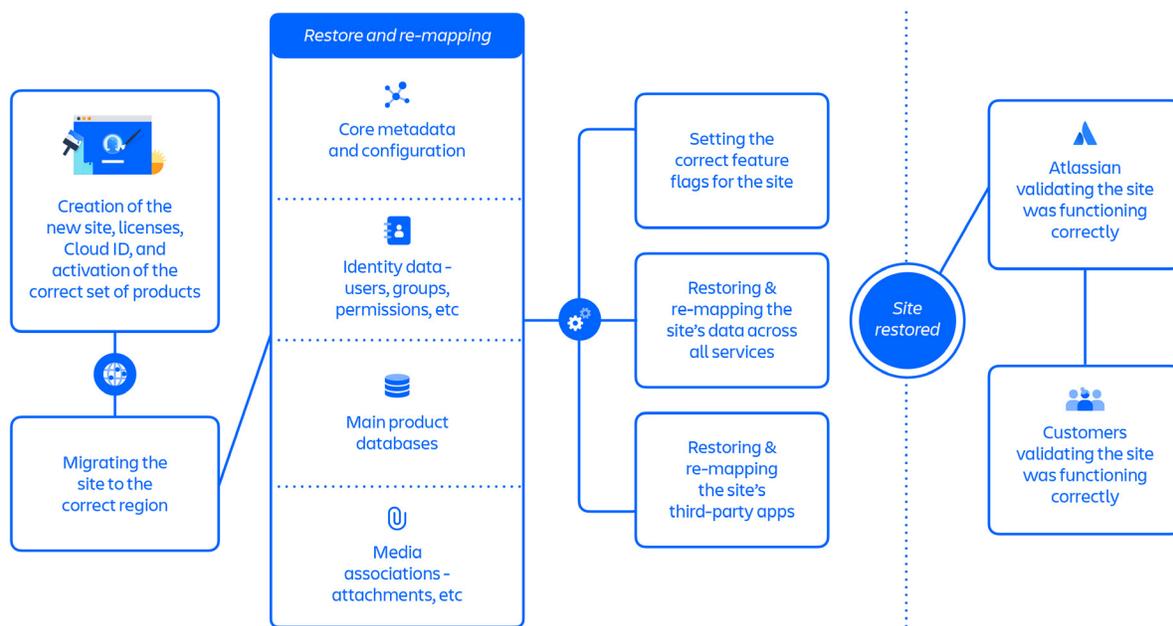


Figura 7: Passaggi chiave nell'approccio Restoration 1.

L'approccio Restoration 1 comprendeva circa 70 passaggi individuali che, se aggregati a livello generale, seguivano un flusso perlopiù sequenziale di:

- Creazione del nuovo sito, licenze, Cloud ID e attivazione del set di prodotti corretto
- Migrazione del sito alla regione corretta
- Ripristino e nuova mappatura della configurazione e dei metadati principali del sito
- Ripristino e nuova mappatura dei dati di identità del sito: utenti, gruppi, autorizzazioni, ecc.
- Ripristino dei principali database di prodotto del sito
- Ripristino a nuova mappatura delle associazioni multimediali del sito: allegati, ecc.
- Impostazione dei corretti flag delle funzioni per il sito
- Ripristino e nuova mappatura dei dati del sito in tutti i servizi
- Ripristino e nuova mappatura delle app di terze parti del sito
- Convalida, da parte di Atlassian, del corretto funzionamento del sito
- Convalida, da parte dei clienti, del corretto funzionamento del sito

Una volta ottimizzato, l'approccio Restoration 1 ha richiesto circa 48 ore per ripristinare un lotto di siti e tra il 5 e il 14 aprile è stato utilizzato per il recupero del 53% degli utenti interessati su 112 siti.

Flusso di lavoro 3: ripristino accelerato e approccio Restoration 2

Timestamp: giorni 4-13 (9-17 aprile)

Con l'approccio Restoration 1, ci sarebbero volute tre settimane per ripristinare tutti i clienti. Pertanto, il 9 aprile abbiamo proposto un nuovo approccio per accelerare il restauro di tutti i siti, Restoration 2 (vedere la *Figura 8* di seguito).

L'approccio Restoration 2 ha offerto una migliore parallelizzazione tra le fasi di ripristino riducendo la complessità e il numero di dipendenze presenti nell'approccio Restoration 1.

L'approccio Restoration 2 ha comportato la ricreazione (o l'annullamento dell'eliminazione) dei record associati al sito in tutti i rispettivi sistemi, a partire dal record del Servizio del Catalogo. Un elemento chiave di questo nuovo approccio è stato il *riutilizzo di tutti i precedenti identificatori di sito*. Ciò ha rimosso dal processo precedente più della metà dei passaggi utilizzati per mappare i vecchi identificatori ai nuovi identificatori, ivi compresa la necessità di coordinarsi con ogni fornitore di app di terze parti per ciascun sito.

Tuttavia, il passaggio dall'approccio Restoration 1 all'approccio Restoration 2 ha aggiunto un notevole sovraccarico alla risposta agli imprevisti:

- Molti degli script e dei processi di automazione implementati nell'approccio Restoration 1 hanno dovuto essere modificati per l'approccio Restoration 2.
- I team che eseguivano ripristini (ivi compresi i coordinatori degli imprevisti) hanno dovuto gestire batch paralleli di ripristini in entrambi gli approcci, mentre il processo di Restoration 2 veniva testato e convalidato.
- L'utilizzo di un nuovo approccio ha implicato la necessità di testare e convalidare il processo di Restoration 2 prima di estenderne l'applicazione, il che ha richiesto la duplicazione del lavoro di convalida precedentemente portato a termine per Restoration 1.

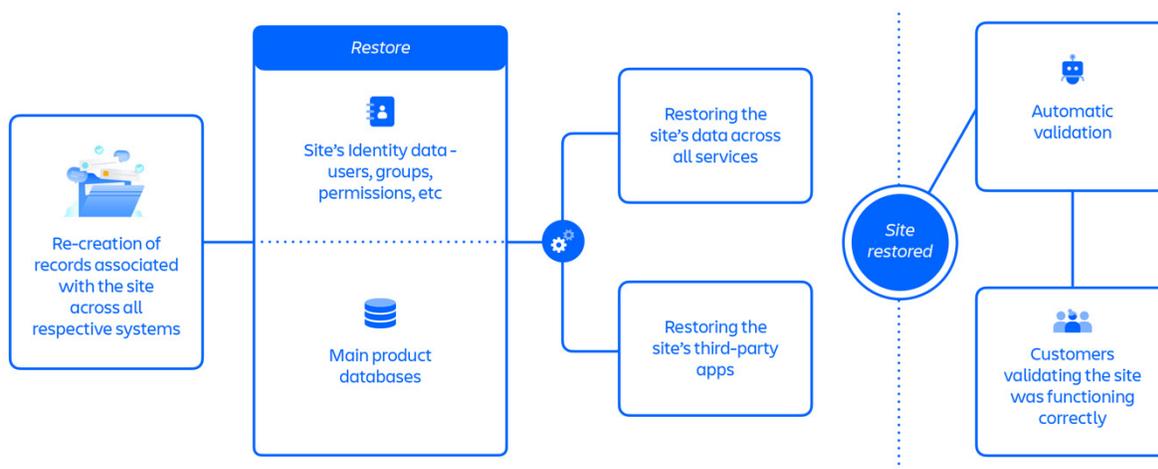


Figura 8: Passaggi chiave nell'approccio Restoration 2.

Il grafico sopra riportato rappresenta l'approccio Restoration 2, che prevedeva oltre 30 passaggi in un flusso in gran parte parallelizzato costituito da:

- Ricreazione di record associati al sito in tutti i rispettivi sistemi
- Ripristino dei dati Identity del sito: utenti, gruppi, autorizzazioni, ecc.
- Ripristino dei principali database di prodotto del sito
- Ripristino dei dati del sito in tutti i servizi
- Ripristino delle app di terze parti del sito
- Convalida automatica
- Convalida, da parte dei clienti, del corretto funzionamento del sito

Nell'ambito del ripristino accelerato, abbiamo anche adottato misure per concentrare nella fase iniziale e automatizzare il ripristino del sito perché per i lotti di grandi dimensioni il ripristino manuale non avrebbe offerto una scalabilità adeguata. La natura sequenziale del processo di ripristino implica la possibilità che il ripristino del sito sia più

lento per il ripristino di database di grandi dimensioni e il ripristino della base utenti/delle autorizzazioni. Le ottimizzazioni da noi implementate includevano quanto segue:

- Abbiamo sviluppato gli strumenti e le regole necessarie per *concentrare nelle fasi iniziali*, e poi implementare nel lungo termine, azioni come il ripristino dei database e la sincronizzazione con Identity, per consentirne il completamento prima di altre fasi del ripristino.
- I team di ingegneri hanno creato l'automazione per le singole fasi rendendo possibile eseguire in sicurezza il ripristino in lotti di grandi dimensioni.
- L'automazione è stata sviluppata per convalidare il corretto funzionamento dei siti una volta completate tutte le fasi di ripristino.

Il ripristino attraverso il più veloce approccio Restoration 2 ha richiesto circa 12 ore ed è stato utilizzato per il recupero del 47% circa degli utenti interessati in 771 siti tra il 14 e il 17 aprile.

Perdita di dati minima in seguito al ripristino dei siti eliminati

Il backup dei nostri database viene eseguito utilizzando una combinazione di backup integrali e backup incrementali che ci consentono di scegliere un particolare momento «temporizzato» per ripristinare i nostri archivi dati durante il periodo di conservazione del backup (30 giorni). Per la maggior parte dei clienti interessati da questo imprevisto, abbiamo identificato i principali archivi di dati per i nostri prodotti e abbiamo deciso di utilizzare un punto di ripristino risalente a cinque minuti prima della cancellazione dei siti come punto di sincronizzazione sicuro. I data store non primari sono stati ripristinati in base allo stesso punto o riproducendo gli eventi registrati. L'utilizzo di un punto di ripristino fisso per gli archivi primari ci ha permesso di ottenere la coerenza dei dati in tutti gli archivi di dati.

Per 57 clienti oggetto di ripristino nelle prime fasi della nostra risposta agli imprevisti, la mancanza di criteri omogenei e il recupero manuale degli snapshot di backup del database hanno portato al ripristino di alcuni database Confluence e Insight a un momento *anteriore* dei cinque minuti prima della cancellazione del sito. L'incoerenza è emersa durante un processo di audit post-ripristino. Abbiamo quindi recuperato il resto dei dati, contattato i clienti interessati e li stiamo aiutando ad applicare modifiche per ripristinare ulteriormente i loro dati.

In sintesi

- Durante questo imprevisto abbiamo raggiunto il nostro Obiettivo punto di ripristino (RPO) di un'ora.

- La perdita di dati in conseguenza dell'imprevisto è limitata a cinque minuti prima della cancellazione del sito.
- Per un numero limitato di clienti il ripristino dei database Confluence o Insight risale a un anteriore ai cinque minuti prima dell'eliminazione del sito. Tuttavia, siamo in grado di recuperare quei dati e stiamo attualmente lavorando con i clienti per ripristinarli.

Comunicazione di imprevisti

Quando parliamo di comunicazioni relative agli imprevisti, include punti di contatto con clienti, partner, media, analisti di settore, investitori e la più ampia community tecnologica.

Cosa è successo

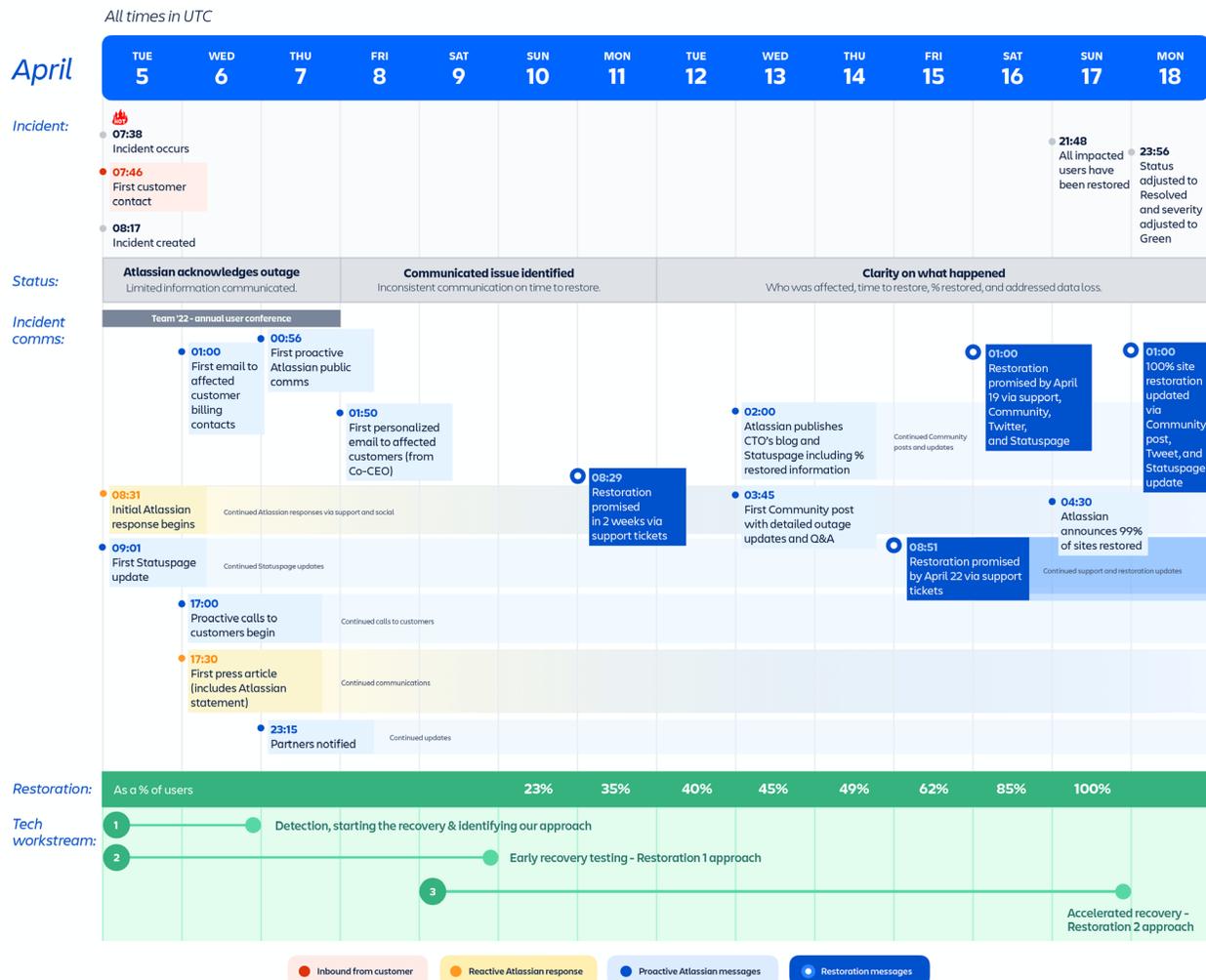


Figura 9: Timeline delle milestone chiave dell'imprevisto.

Timestamp: giorni 1-3 (5-7 aprile)

Prima risposta

La prima richiesta di assistenza è stata inviata il 5 aprile alle 7:46 UTC e l'Assistenza Atlassian ha risposto riconoscendo l'imprevisto entro le 8:31 UTC. Alle 9:03 UTC, è stato pubblicato il primo aggiornamento Statuspage per informare i clienti che stavamo indagando sull'imprevisto. E alle 11:13 UTC, abbiamo confermato tramite Statuspage che avevamo individuato la causa principale e che stavamo lavorando a una correzione. Entro le 1:00 UTC del 6 aprile, nelle comunicazioni iniziali relative alla richiesta del cliente veniva indicato che l'interruzione era dovuta a uno script di manutenzione e che la perdita di dati attesa era minima. Il 6 aprile alle 17:30 UTC Atlassian ha risposto alle domande dei media con una dichiarazione. Il 7 aprile alle 00:56 UTC, Atlassian ha pubblicato su Twitter un primo messaggio esterno ad ampia diffusione, in cui prendeva atto dell'imprevisto.

Timestamp: giorni 4-7 (8-11 aprile)

Primi contatti su ampia scala con i clienti

L'8 aprile alle 1:50 UTC, Atlassian ha inviato ai clienti interessati le scuse del co-fondatore e co-CEO Scott Farquhar. Nei giorni successivi, abbiamo lavorato per ripristinare le informazioni di contatto eliminate e creare richieste di assistenza per tutti i siti interessati che non ne avevano ancora inviata una. Il nostro team di supporto ha quindi continuato a inviare aggiornamenti regolari sulle operazioni di ripristino tramite le richieste di supporto associate a ciascun sito interessato.

Timestamp: giorni 8-14 (12-18 aprile)

Maggiore chiarezza e completamento del ripristino

Il 12 aprile, [Atlassian ha pubblicato un aggiornamento del CTO, Sri Viswanath](#), fornendo maggiori dettagli tecnici su ciò che era successo, chi era stato colpito, se c'era stata perdita di dati, l'avanzamento nel ripristino e l'indicazione che sarebbero potute essere necessarie fino a due settimane per un ripristino completo di tutti i siti. Il blog è stato accompagnato da un'altra dichiarazione stampa a nome di Sri. Abbiamo inoltre fatto riferimento al blog di Sri nel nostro [primo post proattivo sulla Atlassian Community, firmato dal responsabile della progettazione, Stephen Deasy](#), che successivamente è diventato il luogo dedicato per ulteriori aggiornamenti e domande e risposte con il pubblico più in generale. Un aggiornamento del 18 aprile a questo post ha annunciato il ripristino completo di tutti i siti dei clienti interessati.



Perché non abbiamo dato prima una risposta al pubblico?

1. Abbiamo dato priorità alla comunicazione diretta con i clienti interessati tramite Statuspage, e-mail, richieste di supporto e interazioni individuali. Tuttavia, non siamo stati in grado di raggiungere molti clienti perché abbiamo perso i loro recapiti quando i loro siti sono stati cancellati. Avremmo dovuto implementare molto prima comunicazioni di maggiore portata, al fine di informare i clienti e gli utenti finali interessati in merito alla nostra Risposta agli imprevisti e in merito alla timeline di risoluzione.
2. Pur sapendo sin dall'inizio che cosa avesse causato l'imprevisto, la complessità dell'architettura e le particolari circostanze uniche di questo imprevisto hanno rallentato la nostra capacità di valutare rapidamente e stimare con precisione i tempi di risoluzione. Anziché aspettare di avere un quadro completo, avremmo dovuto comunicare con trasparenza ciò che sapevamo e ciò che non sapevamo. Fornire stime generali sui ripristini (anche se approssimative) e dire chiaramente quando ci aspettavamo di avere un quadro più completo avrebbe permesso ai nostri clienti di pianificare meglio come affrontare l'imprevisto. Ciò è vale soprattutto per gli amministratori di sistema e i referenti tecnici, che sono in prima linea nella gestione degli stakeholder e degli utenti all'interno delle loro organizzazioni.

Esperienza di assistenza e comunicazione con i clienti

Come accennato in precedenza, lo stesso script che ha eliminato i siti dei clienti ha eliminato anche gli identificativi chiave dei clienti e le informazioni di contatto (ad es. URL cloud, recapiti dell'amministratore di sistema del sito) dai nostri ambienti di produzione. Si tratta di un aspetto rilevante perché i nostri sistemi principali (ad es. supporto, licenze, fatturazione) si basano tutti sull'esistenza di un URL cloud e sui recapiti dell'amministratore di sistema del sito come identificatori primari a fini di sicurezza, routing e prioritizzazione. Quando abbiamo perso questi identificatori, abbiamo perso in un primo momento la nostra capacità di individuare i clienti e interagire sistematicamente con loro.

In che modo ne ha risentito il supporto per i nostri clienti?

Innanzitutto, la maggior parte dei clienti interessati non è riuscita a comunicare con il nostro team di supporto tramite il normale [modulo di contatto online](#). Questo modulo è progettato in modo tale da richiedere a un utente di accedere con il proprio ID Atlassian e di fornire un URL cloud valido. Senza un URL valido, l'utente non ha la possibilità di inviare una richiesta di supporto tecnico. In condizioni operative normali, questa verifica è intenzionale per garantire la sicurezza del sito e la valutazione delle richieste. Tuttavia, tale requisito ha creato un risultato indesiderato per i clienti interessati dall'interruzione perché ha impedito loro di inviare una richiesta di supporto per il sito ad alta priorità.

In secondo luogo, l'eliminazione dei dati dell'amministratore di sistema del sito causata dall'imprevisto ha posto in essere una lacuna nella nostra capacità di interagire in modo proattivo con i clienti interessati. Nei primi giorni dell'imprevisto, abbiamo inviato comunicazioni proattive ai referenti tecnici e di fatturazione del cliente interessati registrati presso Atlassian. Tuttavia, abbiamo subito scoperto che molti recapiti tecnici e di fatturazione dei clienti interessati erano obsoleti. Senza le informazioni dell'amministratore di sistema per ogni sito, non disponevamo di un elenco completo di recapiti attivi e autorizzati attraverso i quali interagire.

Come abbiamo risposto?

I nostri team di supporto avevano tre priorità altrettanto importanti per accelerare il ripristino del sito e porre rimedio all'interruzione dei nostri canali di comunicazione nei primi giorni dell'imprevisto.

Innanzitutto, ottenere un elenco attendibile di recapiti convalidati per i clienti. Mentre i nostri team di ingegneri lavoravano per ripristinare i siti dei clienti, i nostri team a contatto con i clienti si sono concentrati sul ripristino di informazioni di contatto convalidate. Abbiamo utilizzato ogni meccanismo a nostra disposizione (sistemi di fatturazione, richieste di assistenza preventiva, altri backup protetti degli utenti, comunicazione diretta con i clienti, ecc.) per ricostruire la lista di contatti. Il nostro obiettivo era quello di avere una richiesta di supporto relativa agli imprevisti per ogni sito interessato onde semplificare il contatto diretto e i tempi di risposta.

In secondo luogo, ristabilire flussi di lavoro, code e SLA specifici per questo imprevisto. L'eliminazione del Cloud ID e l'impossibilità di autenticare correttamente gli utenti hanno influito anche sulla nostra capacità di elaborare le richieste di assistenza relative all'imprevisto servendoci dei nostri normali sistemi. Le richieste non venivano visualizzate correttamente nelle code e nelle dashboard relative a priorità ed escalation. Abbiamo rapidamente creato un team interfunzionale (supporto, prodotto, IT) per progettare e aggiungere ulteriore logica, SLA, stato dei flussi di lavoro e dashboard. Siccome tutto

ciò doveva essere fatto all'interno del nostro sistema di produzione, ci sono voluti diversi giorni per portarne a termine lo sviluppo, i test e la distribuzione.

In terzo luogo, aumentare in modo massiccio le convalide manuali per accelerare il ripristino dei siti. Man mano che la funzione di ingegneria avanzava con i ripristini iniziali, è diventato chiaro che sarebbe stata necessario ricorrere alla capacità dei nostri team di supporto globali per contribuire ad accelerare il ripristino dei siti tramite test manuali e controlli di convalida. Tale processo di convalida sarebbe diventato un percorso critico per far arrivare i siti ripristinati ai nostri clienti, una volta che il team di ingegneria avesse accelerato il ripristino dei dati. Abbiamo dovuto creare un flusso indipendente di procedure operative standard (SOP), flussi di lavoro, passaggi di consegne ed elenchi di personale, mobilitando oltre 450 tecnici del supporto per eseguire controlli di convalida, implementando turni per fornire una copertura 24 ore su 24, 7 giorni su 7 e accelerare il ripristino a beneficio dei clienti.

Pur avendo definito chiaramente queste priorità chiave entro la fine della prima settimana, eravamo limitati nella nostra capacità di fornire aggiornamenti *significativi* per la poca visibilità delle tempistiche di risoluzione degli imprevisti, dovuta alla complessità dei processi di ripristino. Avremmo dovuto prendere atto più velocemente della nostra incertezza nel fornire una data di ripristino del sito e renderci disponibili più velocemente a un dialogo di persona, per consentire ai nostri clienti di pianificare di conseguenza.

Cosa miglioreremo?

Abbiamo immediatamente bloccato le eliminazioni in blocco dai siti fino a quando non saranno state apportate le modifiche appropriate.

Lasciandoci alle spalle questo imprevisto e rivalutando i nostri processi interni, vogliamo prendere atto del fatto che non sono le persone a causare imprevisti. Sono piuttosto i sistemi a consentire di commettere errori. In questa sezione vengono riassunti i fattori che hanno contribuito a questo imprevisto. In questa sede parliamo anche dei nostri piani per accelerare la correzione di queste carenze e questi problemi.

Insegnamento 1: le «eliminazioni temporanee» devono essere universali su tutti i sistemi

In generale, un'eliminazione di questo tipo dovrebbe essere vietata o avere più livelli di protezione per evitare errori. Il miglioramento principale in corso di attuazione è di impedire a livello globale l'eliminazione dei dati e metadati dei clienti che non sono stati sottoposti a un processo di eliminazione temporanea.

a) L'eliminazione dei dati deve avvenire solo come eliminazione temporanea

L'eliminazione di un intero sito deve essere vietata e l'eliminazione temporanea deve richiedere protezioni multilivello per prevenire errori. Implementeremo una politica di «eliminazione temporanea», impedendo a script o sistemi esterni di eliminare i dati dei clienti in un ambiente di produzione. La nostra politica di «eliminazione temporanea» consentirà una conservazione dei dati sufficiente a consentire il ripristino rapido e sicuro dei dati. I dati verranno eliminati dall'ambiente di produzione solo alla scadenza di un periodo di conservazione.

Azioni

- ✓ **Implementazione di una «eliminazione temporanea» nei flussi di lavoro di provisioning e in tutti gli archivi dati pertinenti:** inoltre, il team di Tenant Platform verificherà che l'eliminazione dei dati possa avvenire solo dopo le disattivazioni, così come altre misure di sicurezza in questo spazio. Nel lungo termine, Tenant Platform assumerà un ruolo di primo piano per sviluppare ulteriormente una corretta gestione statale dei dati tenant.

b) L'eliminazione temporanea deve disporre di un processo di esame standardizzato e verificato

Le azioni di eliminazione temporanea sono operazioni ad alto rischio. Pertanto, dobbiamo disporre di processi di revisione standardizzati o automatizzati che includano rollback e procedure di test definiti per affrontare simili operazioni.

Azioni

- ✓ **Implementazione graduale forzata di qualsiasi azione di eliminazione temporanea:** tutte le nuove operazioni che richiedono l'eliminazione vengono prima testate all'interno dei nostri siti per convalidare l'approccio usato e verificare l'automazione. Una volta completata la convalida, sposteremo progressivamente i clienti attraverso lo stesso processo e continueremo a testare le irregolarità prima di applicare l'automazione all'intera base di utenti selezionata.
- ✓ **Le azioni di eliminazione temporanea devono avere un piano di rollback testato:** qualsiasi attività di eliminazione graduale dei dati deve testare il ripristino dei dati eliminati prima di essere eseguita in produzione e disporre di un piano di rollback testato.

Insegnamento 2: nell'ambito del programma DR, automatizzare il ripristino per eventi di eliminazione multi-sito e multi-prodotto per un gruppo più ampio di clienti

[Atlassian Data Management](#) descrive in dettaglio i nostri processi di gestione dei dati. Per garantire un'elevata disponibilità, eseguiamo il provisioning e manteniamo una replica di standby sincrona in più zone di disponibilità (AZ) AWS. Il failover AZ è automatizzato e richiede in genere 60-120 secondi e gestiamo regolarmente le interruzioni nel data center e altre interruzioni comuni senza alcun impatto sui clienti.

Manteniamo anche backup immutabili progettati per resistere agli eventi di corruzione dei dati, che consentono il ripristino a un punto nel tempo precedente. I backup vengono conservati per 30 giorni e Atlassian testa e verifica continuamente i backup di archiviazione per il ripristino. Se necessario, possiamo ripristinare tutti i clienti contemporaneamente in un nuovo ambiente.

Utilizzando questi backup, eseguiamo regolarmente il rollback di singoli clienti o di un piccolo gruppo di clienti che eliminano per errore i propri dati. Tuttavia, la cancellazione a livello di sito non aveva runbook che potessero essere rapidamente automatizzati per la portata di questo evento che richiedeva il ricorso coordinato a strumenti e automazione in tutti i prodotti e servizi.

Ciò che non abbiamo (ancora) automatizzato è il ripristino di un ampio sottoinsieme di clienti nel nostro ambiente esistente (e attualmente in uso) senza avere un impatto su nessuno dei nostri altri clienti.

All'interno del nostro ambiente cloud, ogni archivio di dati contiene dati di più clienti. Poiché i dati eliminati in questo imprevisto costituivano solo una parte degli archivi di dati che continuano a essere utilizzati da altri clienti, dobbiamo estrarre e ripristinare manualmente i singoli dati dai nostri backup. Il ripristino di ogni sito dei clienti è un processo lungo e complesso, che richiede la convalida interna e la verifica finale del cliente quando il sito viene ripristinato.

Azioni



Accelerazione dei ripristini multi-prodotto e multi-sito per un gruppo più ampio di clienti: il programma di DR soddisfa i nostri attuali standard RPO di un'ora. Sfrutteremo l'automazione e le conoscenze acquisite da questo imprevisto per accelerare il programma di DR e soddisfare l'RTO definito nella nostra politica per imprevisti di questa portata.



Automazione e aggiunta della verifica di questo caso al test di DR:

eseguiremo regolarmente procedure di DR che prevedono il ripristino di tutti i prodotti per grandi gruppi di siti. Questi test di DR verificheranno che i runbook siano aggiornati man mano che la nostra architettura si evolve e si riscontrano nuovi casi limite. Miglioreremo continuamente il nostro approccio al restauro, automatizzeremo una parte maggiore del processo di ripristino e ridurremo i tempi dello stesso.

Insegnamento 3: migliorare il processo di gestione degli imprevisti per eventi su larga scala

Il nostro programma di gestione degli imprevisti è adatto per la gestione degli incidenti di maggiore e minore entità che si sono verificati nel corso degli anni. Spesso simuliamo la risposta agli imprevisti per imprevisti di scala minore e di durata più breve, che in genere coinvolgono meno persone e team.

Tuttavia, al suo apice, questo imprevisto ha portato al coinvolgimento di centinaia di ingegneri e dipendenti dell'assistenza clienti che hanno lavorato contemporaneamente per ripristinare i siti dei clienti. Il nostro programma di gestione degli imprevisti e i nostri team non erano strutturati per gestire la profondità, l'ampiezza e la durata di questo tipo di imprevisto (vedere la *Figura 10* riportata di seguito).

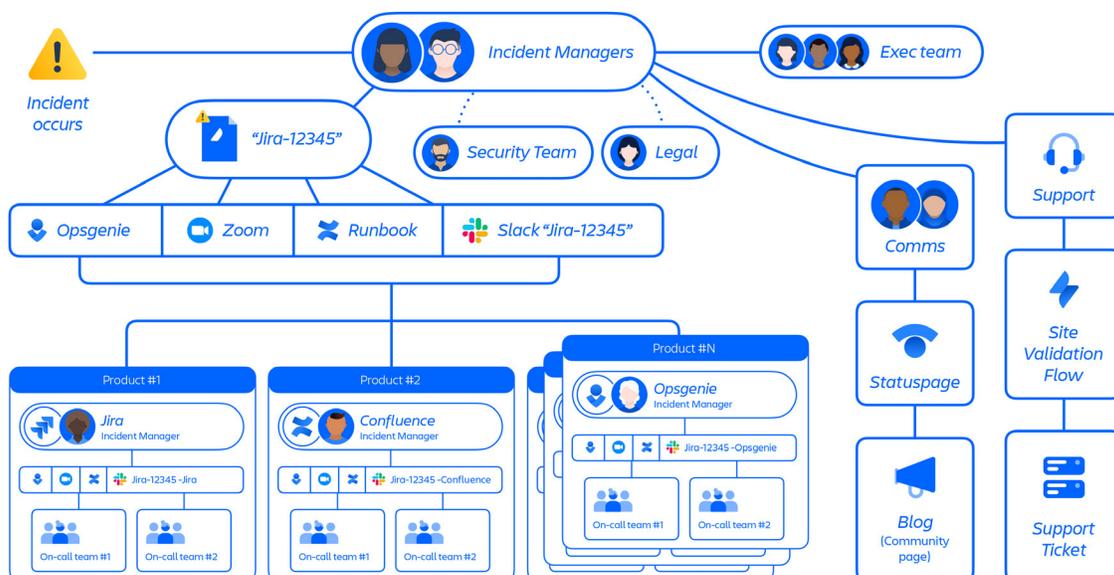


Figura 10: panoramica del processo di gestione degli imprevisti su larga scala.

Il nostro processo di gestione degli imprevisti su larga scala sarà definito meglio e praticato spesso

Disponiamo di guide operative per imprevisti a livello di prodotto, ma non per eventi di questa portata, con centinaia di persone che lavorano contemporaneamente in tutta l'azienda. Gli strumenti di gestione degli imprevisti presentano un'automazione che crea flussi di comunicazione come Slack, Zoom e documenti Confluence, ma non crea i sottoflussi necessari per isolare i flussi di ripristino in caso di imprevisti su larga scala.

Azioni



Definizione di una guida operativa e di strumenti per imprevisti su larga scala e conduzione di procedure simulate: definire e documentare i tipi di imprevisto che possono essere considerati su larga scala e richiedono questo livello di risposta. Descrivere le fasi di coordinamento chiave e sviluppare strumenti che aiutino i Gestori Imprevisti e altre funzioni aziendali a semplificare la risposta e avviare il ripristino. I Gestori Imprevisti e i loro team faranno regolarmente simulazioni, corsi di formazione e azioni di perfezionamento di strumenti e documenti per migliorare continuamente.

Insegnamento 4: migliorare i nostri processi di comunicazione

a) Abbiamo eliminato gli identificativi critici dei clienti, influenzando le comunicazioni e le azioni per i soggetti interessati

Lo stesso script che ha eliminato i siti dei clienti ha eliminato anche gli identificativi chiave dei clienti (ad es. URL del sito, recapiti dell'amministratore di sistema del sito) dai nostri ambienti di produzione. Di conseguenza, (1) ai clienti è stato impedito di presentare richieste di supporto tecnico tramite il nostro normale canale di supporto; (2) ci sono voluti diversi giorni per ottenere un elenco attendibile dei recapiti dei clienti chiave (come gli amministratori di sistema del sito) interessati dall'interruzione per un coinvolgimento proattivo; e (3) i flussi di lavoro di supporto, gli SLA, le dashboard e i processi di escalation inizialmente non funzionavano correttamente a causa della particolare natura dell'imprevisto.

Durante l'interruzione, le escalation dei clienti sono arrivate anche attraverso più canali (e-mail, telefonate, richieste del CEO, LinkedIn e altri canali social e richieste di supporto). Strumenti e processi disparati tra i nostri team a contatto con i clienti hanno rallentato la nostra risposta e reso più difficile il monitoraggio e la segnalazione olistici di queste escalation.

b) Non avevamo una guida operativa sulla comunicazione degli imprevisti sufficientemente completa per affrontare questo livello di complessità

Non disponevamo di una guida operativa sulla comunicazione degli imprevisti che descrivesse principi, ruoli e responsabilità per mobilitare abbastanza rapidamente un team unificato e interfunzionale per la comunicazione degli imprevisti. Non abbiamo reso possibile il riconoscimento degli imprevisti in modo rapido e coerente attraverso più canali, in particolare sui social media. L'approccio corretto sarebbe stato quello di effettuare comunicazioni pubbliche di più ampia portata sull'interruzione, nonché ripetere l'importantissimo messaggio che non si è verificata alcuna perdita di dati e che l'imprevisto non era il risultato di un attacco informatico.

Azioni

- ✓ **Miglioramento del backup dei contatti chiave:** eseguire il backup delle informazioni di contatto dell'account autorizzate, all'esterno dell'istanza del prodotto.
- ✓ **Strumenti di supporto per il retrofit:** creare meccanismi per i clienti senza un URL del sito valido o un ID Atlassian per entrare in contatto diretto con il nostro team di supporto tecnico.
- ✓ **Sistema e processi di escalation dei clienti:** investire in un sistema di escalation e flussi di lavoro unificati e basati su account che consentano di archiviare più oggetti di lavoro (richieste, task, ecc.) sotto un singolo oggetto account cliente, per migliorare coordinamento e visibilità in tutti i nostri team a contatto con i clienti.
- ✓ **Accelerare la copertura 24 ore su 24, 7 giorni su 7, della gestione delle escalation:** eseguire piani di espansione dell'impronta globale per la funzione di gestione delle escalation per consentire una copertura costante 24 ore su 24, 7 giorni su 7, con personale designato con sede in tutte le principali regioni geografiche insieme a ruoli di supporto per fornire a esperti e leader l'assistenza richiesta in materia di prodotti e vendite.
- ✓ **Aggiornamento della nostra guida operativa per le comunicazioni sugli imprevisti con nuove lezioni e regolare rivisitazione:** rivisitare la guida operativa per definire internamente chiari ruoli e linee di comunicazione. Utilizzare il framework [DACI](#) per gli imprevisti e disporre di back-up 24 ore su 24, 7 giorni su 7, per ogni ruolo in caso di malattia, ferie o altri eventi imprevisti. Effettuare un audit trimestrale per verificare la disponibilità in ogni momento.

Azioni (segue)

Seguire il modello per la comunicazione degli imprevisti in tutte le comunicazioni: spiegare cosa è successo, chi è stato interessato, timeline per il ripristino, percentuali di ripristino del sito, perdita di dati prevista, con i livelli di affidabilità associati, insieme a indicazioni chiare su come contattare il supporto.

Osservazioni conclusive

Mentre l'interruzione viene risolta e i clienti usufruiscono di un completo ripristino, il nostro lavoro continua. In questa fase, stiamo implementando le modifiche descritte sopra per migliorare i nostri processi, aumentare la nostra resilienza ed evitare il ripetersi di una situazione come questa.

Atlassian è un'organizzazione orientata all'apprendimento e i nostri team hanno sicuramente imparato molte dure lezioni da questa esperienza. Stiamo mettendo in pratica queste lezioni per apportare cambiamenti duraturi al nostro business. In definitiva, grazie a questa esperienza emergeremo più forti e forniremo un servizio migliore.

Ci auguriamo che gli insegnamenti tratti da questo imprevisto siano utili ad altri team che stanno lavorando diligentemente per fornire servizi affidabili ai propri clienti.

Infine, voglio ringraziare coloro che stanno leggendo questo articolo e stanno imparando con noi e coloro che fanno parte del team e della Atlassian Community.

-Sri Viswanath, CTO