ATLASSIAN

Nachträgliche Vorfallanalyse

Ausfall April 2022

Diese Übersetzung wird nur zum besseren Verständnis bereitgestellt. Im Falle von Unklarheiten oder Konflikten

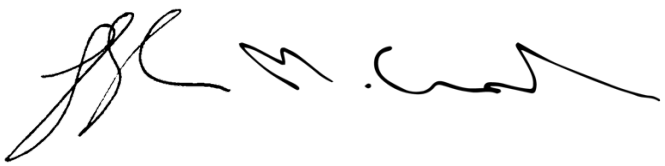
Schreiben von unseren Mitbegründern und Co-CEOs

Wir möchten den Ausfall bestätigen, der Anfang dieses Monats den Service für unsere Kunden unterbrochen hat. Es ist uns bewusst, dass unsere Produkte für Ihr Unternehmen von entscheidender Bedeutung sind, und wir nehmen diese Verantwortung nicht auf die leichte Schulter. Versagt unser Service, kommt dies Ihr Unternehmen teuer zu stehen. Mehr gibt es dazu nicht zu sagen. Wir arbeiten daran, das Vertrauen der betroffenen Kunden zurückzugewinnen.

> Einer unserer Kernwerte bei Atlassian ist eine "Offene Unternehmenskultur – kein Bullsh**". Wir erwecken diesen Wert zum Leben, indem wir beispielsweise Vorfälle offen diskutieren und sie als Chance nutzen, etwas Neues zu lernen. Wir veröffentlichen diese nachträgliche Vorfallanalyse für unsere Kunden, unsere Atlassian Community und die breitere technische Community. Atlassian ist stolz auf seinen Vorfallmanagementprozess, der betont, dass eine Kultur ohne Schuldzuweisungen sowie der Fokus auf Verbesserungsmöglichkeiten unserer technischen Systeme und Prozesse entscheidend für die Bereitstellung hochwertiger, vertrauenswürdiger Services ist. Wir geben unser Bestes, um jegliche Art von Vorfällen zu vermeiden, sind aber auch der Meinung, dass Vorfälle ein wirksames Mittel zur Verbesserung sind.

Wir versichern Ihnen, dass die Cloud-Plattform von Atlassian die vielfältigen Anforderungen unserer über 200.000 Cloud-Kunden jeder Größe und Branche erfüllt. Vor diesem Vorfall hat unsere Cloud durchwegs eine Verfügbarkeit von 99,9 % erreicht und die SLAs für die Verfügbarkeit übertroffen. Wir haben langfristige Investitionen in unsere Plattform und in eine Reihe zentralisierter Plattformfunktionen getätigt, mit einer skalierbaren Infrastruktur und stetigen Verbesserungen bei der Sicherheit.

Wir danken Ihnen, unseren Kunden und Partnern, für Ihr anhaltendes Vertrauen und die Zusammenarbeit. Wir hoffen, dass die in diesem Dokument beschriebenen Details und Maßnahmen zeigen, dass Atlassian auch weiterhin eine erstklassige Cloud-Plattform und ein leistungsstarkes Produktportfolio bereitstellen wird, um den Ansprüchen aller Teams gerecht zu werden.



– Scott und Mike

Zusammenfassung

Am Dienstag, den 5. April 2022, verloren 775 Atlassian-Kunden ab 7:38 Uhr (UTC) den Zugang zu ihren Atlassian-Produkten. Für einige dieser Kunden dauerte der Ausfall bis zu 14 Tage. Bis zum 8. April konnten wir den Ausfall für die ersten Kunden beheben und am 18. April hatten alle Kunden wieder Zugriff auf unseren Service.

Der Ausfall war nicht das Ergebnis eines Cyberangriffs und es gab keinen unbefugten Zugriff auf Kundendaten. Atlassian verfügt über ein umfassendes Datenverwaltungsprogramm mit veröffentlichten SLAs. Bisher konnten wir diese SLAs stets übertreffen.

Obwohl es sich um einen schwerwiegenden Vorfall handelte, verlor kein Kunde mehr als fünf Minuten an Daten. Darüber hinaus nutzten über 99,6 % unserer Kunden und Benutzer unsere Cloud-Produkte während der Wiederherstellung ohne weitere Unterbrechung.

> In diesem Dokument bezeichnen wir Kunden, deren Sites im Rahmen dieses Vorfalls gelöscht wurden, als "betroffene" Kunden. Diese nachträgliche Vorfallanalyse liefert die genauen Details des Vorfalls, beschreibt die Schritte, die wir zur Behebung ergriffen haben, und wie wir in Zukunft Situationen wie diese vermeiden möchten. In diesem Abschnitt finden Sie eine allgemeine Zusammenfassung des Vorfalls, mit weiteren Einzelheiten im Rest des Dokuments.

Was ist geschehen?

2021 haben wir die Übernahme und Integration einer eigenständigen Atlassian-App für Jira Service Management und Jira Software namens "Insight – Asset Management" abgeschlossen. Die Funktionalität dieser eigenständigen App war dann nativ in Jira Service Management und für Jira Software nicht mehr verfügbar. Deshalb mussten wir die eigenständige Legacy-App auf Kunden-Sites löschen, auf denen sie installiert war. Unsere Entwicklungsteams verwendeten ein vorhandenes Skript und einen Prozess, um Instanzen dieser eigenständigen Anwendung zu löschen, es gab jedoch zwei Probleme:

- **Kommunikationsfehler.** Die Kommunikation zwischen dem Team, das die Löschung veranlasst hatte, und dem Team, das mit der Durchführung der Löschung betraut war, war unzureichend. Statt die IDs der *zu löschenden Apps* anzugeben,

stellte das Team die IDs der *gesamten Cloud-Site* bereit, auf der die Apps gelöscht werden sollten.

- **Ungenügende Systemwarnungen.** Die API, die zum Löschen verwendet wurde, akzeptierte sowohl Site- als auch App-Kennungen und ging davon aus, dass die Eingabe korrekt war. Dies bedeutete, dass bei Übergabe einer Site-ID eine Site gelöscht wurde. Wenn eine App-ID übergeben wurde, wurde eine App gelöscht. Es gab kein Warnsignal, um die Art der angeforderten Löschung (Site oder App) zu bestätigen.

Das ausgeführte Skript folgte unserem standardmäßigen Peer-Review-Prozess, der sich darauf konzentrierte, welcher Endpunkt wie aufgerufen wurde. Die bereitgestellten Cloud-Site-IDs wurden nicht gegengeprüft, um festzustellen, ob sie sich auf die Insight App oder auf die gesamte Site bezogen. Das Problem bestand darin, dass das Skript die ID für die gesamte Site eines Kunden enthielt. Das Ergebnis war eine sofortige Löschung von 883 Sites. Dies betraf zwischen 7:38 Uhr (UTC) und 8:01 Uhr (UTC) am Dienstag, den 5. April 2022 775 Kunden. *Siehe "Was ist passiert?"*

Wie haben wir reagiert?

Nachdem der Vorfall am 5. April um 8:17 Uhr (UTC) bestätigt wurde, lösten wir unseren Vorfallmanagementprozess zur Verwaltung schwerwiegender Vorfälle aus und beriefen ein funktionsübergreifendes Vorfallmanagementteam ein. Das globale Team für Incident Response arbeitete für die Dauer des Vorfalls rund um die Uhr, bis alle Sites wiederhergestellt, validiert und an die Kunden zurückgegeben wurden. Darüber hinaus trafen sich die Leiter des Vorfallmanagements alle drei Stunden, um die Arbeitsabläufe zu koordinieren.

Schon früh erkannten wir, wie schwierig es war, den Service für Hunderte Kunden mit verschiedenen Produkten gleichzeitig wiederherzustellen.

Zu Beginn des Vorfalls wussten wir bereits genau, welche Sites betroffen waren, und die Information der genehmigten Eigentümer jeder betroffenen Site über den Ausfall hatte oberste Priorität.

Einige Kontaktdaten von Kunden wurden jedoch gelöscht, weshalb sie keine Support-Tickets einreichen konnten. Dies bedeutete auch, dass wir keinen sofortigen Zugang zu wichtigen Kundenkontakten hatten. *Weitere Einzelheiten finden Sie unter "Allgemeiner Überblick über Wiederherstellungs-Workstreams"*

Wie können wir derartige Situationen in Zukunft vermeiden?

Wir haben eine Reihe von Sofortmaßnahmen ergriffen und verpflichten uns, Änderungen vorzunehmen, um diese Situation in Zukunft zu vermeiden. Nachfolgend sehen Sie vier spezifische Bereiche, in denen wir wesentliche Änderungen vorgenommen haben oder vornehmen werden:

1. **Einrichtung universeller "Vorläufiger Löschung" für alle Systeme**. Eine Löschung in diesem Umfang sollte verboten werden oder mehrere Sicherheitsebenen haben, um Fehler zu vermeiden, darunter einen phasenweisen Rollback-Plan für die "Vorläufige Löschung". Wir werden weltweit die Löschung von Kundendaten und Metadaten verhindern, für die zuvor keine vorläufige Löschung erfolgt ist.

2. **Investition in unser Disaster-Recovery-Programm (DR) für die Automatisierung der Wiederherstellung bei Löschvorgängen, die mehrere Sites und Produkte vieler Kunden betreffen.** Wir werden die Automatisierung und die Erkenntnisse aus diesem Vorfall nutzen, um das DR-Programm zu beschleunigen und das Wiederherstellungszeitziel (RTO) zu erreichen, das in unserer Richtlinie für Vorfälle dieser Größenordnung definiert ist. Wir werden regelmäßig DR-Übungen durchführen, bei denen alle Produkte für diverse Sites wiederhergestellt werden.

3. **Verbesserung des Vorfallmanagements bei schwerwiegenden Vorfällen**. Wir werden unsere Standardanweisungen für schwerwiegende Vorfälle verbessern und sie mit Simulationen dieser Größenordnung üben. Wir werden unsere Schulungen und Tools updaten, um die große Anzahl von parallel arbeitenden Teams zu bewältigen.

4. **Erstellung eines umfangreichen Playbooks für die Kommunikation**. Wir werden Vorfälle frühzeitig über mehrere Kanäle bestätigen. Wir werden innerhalb weniger Stunden öffentliche Mitteilungen zu Vorfällen veröffentlichen. Um die betroffenen Kunden besser zu erreichen, werden wir die Sicherung wichtiger Kontakte verbessern und Support-Tools nachrüsten, damit Kunden ohne gültige URL oder Atlassian-ID direkten Kontakt mit unserem technischen Support-Team aufnehmen können.

Die vollständige Liste der Maßnahmen finden Sie nachfolgend in der nachträglichen Vorfallanalyse. *Siehe "Wie werden wir uns verbessern?"*

Inhaltsverzeichnis

Überblick über die Cloud-Architektur von Atlassian

Um die in diesem Dokument erläuterten Faktoren zu dem Vorfall zu verstehen, ist es hilfreich, zunächst die Bereitstellungsarchitektur für die Produkte, Services und Infrastruktur von Atlassian zu verstehen.

Cloud-Hosting-Architektur von Atlassian

Atlassian nutzt Amazon Web Services (AWS) als Cloud-Serviceanbieter und seine hochverfügbaren Rechenzentrumseinrichtungen in mehreren Regionen weltweit. Jede AWS-Region ist ein separater geografischer Standort mit mehreren, isolierten und physisch getrennten Gruppen von Rechenzentren, die als Availability Zones (AZs) bekannt sind.

Wir nutzen die Rechen-, Speicher-, Netzwerk- und Datendienste von AWS, um unsere Produkte und Plattformkomponenten zu entwickeln, wodurch wir die von AWS angebotenen Redundanzfunktionen wie Availability Zones und Regionen nutzen können.

Architektur für verteilte Services

Mit dieser AWS-Architektur hosten wir eine Reihe von Plattform- und Produktservices, die in unseren Lösungen zum Einsatz kommen. Dazu gehören Plattformfunktionen, die von mehreren Atlassian-Produkten wie Media, Identity, Commerce und unserem Editor gemeinsam genutzt werden, sowie produktspezifische Funktionen wie Jira-Issue-Service und Confluence Analytics.

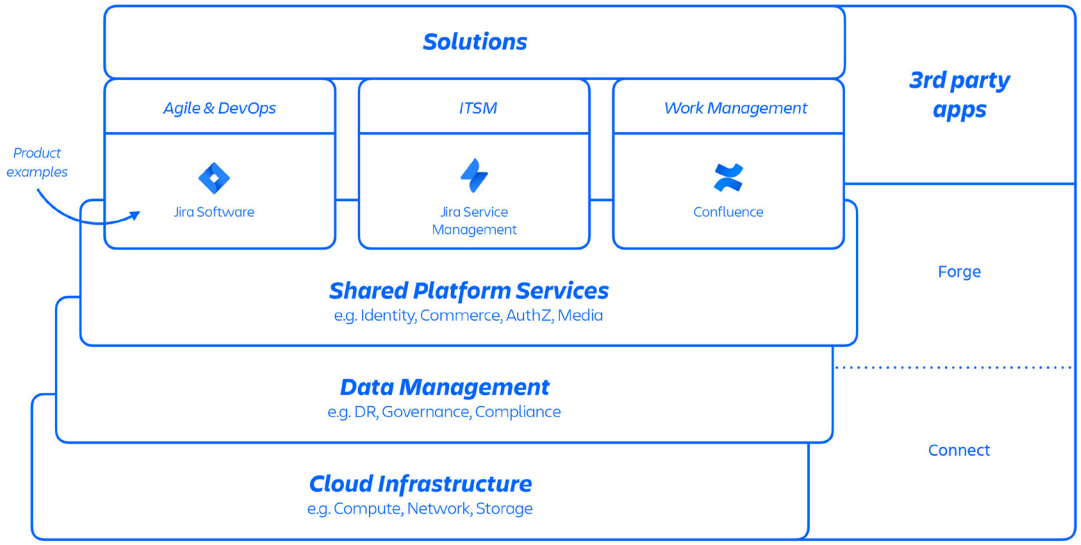


Abbildung 1: Atlassian-Plattformarchitektur

Atlassian-Entwickler stellen diese Services über eine intern entwickelte Platform as a Service (PaaS) namens Micros bereit, die automatisch die Bereitstellung von Shared Services, Infrastruktur, Datenspeichern und deren Verwaltungsfunktionen, einschließlich Sicherheits- und Compliance-Kontrolle, orchestriert (siehe *Abbildung 1* oben). Typischerweise besteht ein Atlassian-Produkt aus mehreren "containerisierten" Services, die mithilfe von Micros auf AWS bereitgestellt werden. Atlassian-Produkte verwenden Kernplattformfunktionen (siehe *Abbildung 2* unten), die vom Anforderungsrouting bis hin zu binären Objektspeichern, Authentifizierung/Autorisierung, transaktionalen benutzergenerierten Inhalten (UGC) und Entitätsbeziehungsspeichern, Data Lakes, allgemeiner Protokollierung, Anforderungsablaufverfolgung, Beobachtbarkeits- und Analyseservices reichen. Diese Microservices basieren auf genehmigten technischen Stacks, die auf Plattformebene standardisiert sind:

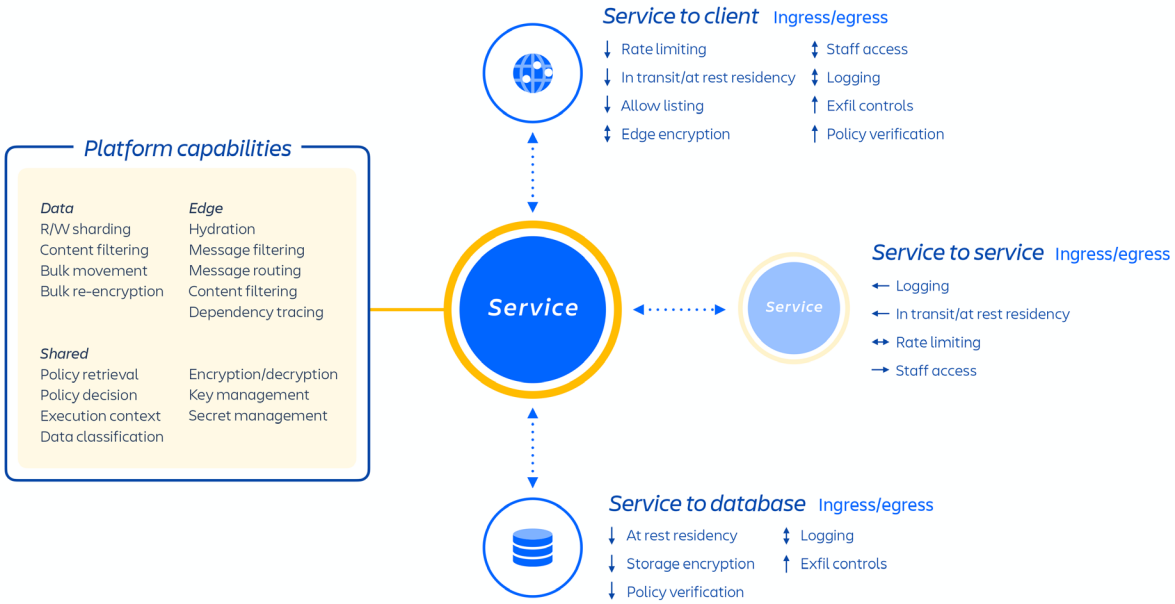


Abbildung 2: Überblick über Atlassian-Microservices

Mehrmandantenarchitektur

Zusätzlich zu unserer Cloud-Infrastruktur haben wir eine mehrmandantenfähige Microservice-Architektur zusammen mit einer geteilten Plattform, die unsere Produkte unterstützt, aufgebaut und betrieben. In einer Mehrmandanten-Architektur bedient ein einziger Service mehrere Kunden, einschließlich Datenbanken und Recheninstanzen, die für die Ausführung unserer Cloud-Produkte erforderlich sind. Jeder Shard (im Wesentlichen ein Container – siehe *Abbildung 3* unten) enthält die Daten für mehrere Mandanten, aber die Daten jedes Mandanten sind isoliert und für andere Mandanten nicht zugänglich.

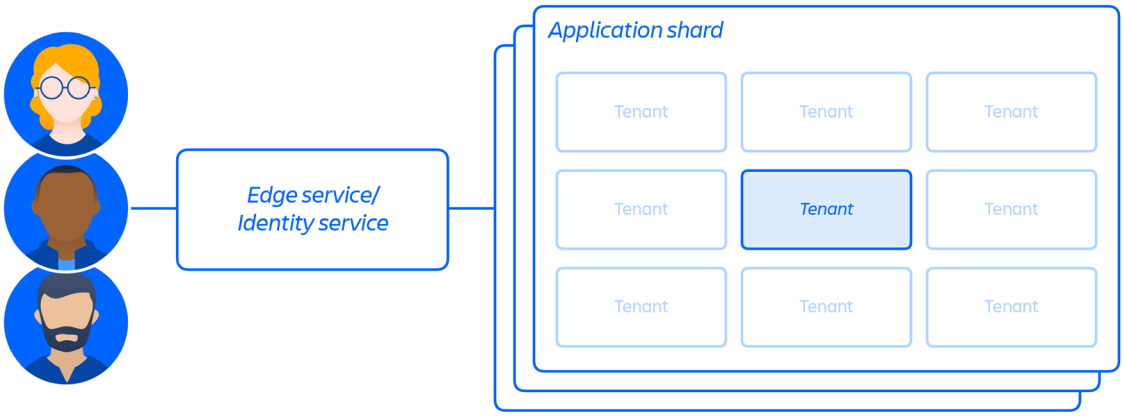


Abbildung 3: Speicherung von Daten in einer Mehrmandantenarchitektur

Bereitstellung und Lebenszyklus von Mandanten

Wenn ein neuer Kunde bereitgestellt wird, löst eine Reihe von Ereignissen die Orchestrierung verteilter Services und die Bereitstellung von Datenspeichern aus. Diese Ereignisse können im Allgemeinen einem von sieben Schritten im Lebenszyklus zugeordnet werden:

1. Commerce-Systeme werden sofort mit den neuesten Metadaten und Zugriffskontrollinformationen für diesen Kunden aktualisiert. Anschließend richtet ein Bereitstellungs-Orchestrierungssystem den "Status der bereitgestellten Ressourcen" durch eine Reihe von Mandanten- und Produktereignissen auf den Lizenzstatus aus.

Mandantenereignisse
Diese Ereignisse betreffen den gesamten Mandanten. Sie können sich auf eines der folgenden Szenarien beziehen:

- Erstellung: ein Mandant wird erstellt und für brandneue Sites verwendet
- Zerstörung: ein ganzer Mandant wird gelöscht

Produkt-Events

- Aktivierung: nach der Aktivierung von lizenzierten Produkten oder Apps von Drittanbietern
- Deaktivierung: nach der Deaktivierung bestimmter Produkte oder Apps
- Aussetzung: nach der Aussetzung eines bestimmten vorhandenen Produkts, wodurch der Zugriff auf eine bestimmte Site, die sie besitzen, deaktiviert wird
- Aufheben der Aussetzung: nach der Aufhebung der Aussetzung eines bestimmten vorhandenen Produkts, wodurch der Zugriff auf eine Site ermöglicht wird, die sie besitzen

Lizenzaktualisierung: enthält Informationen zur Anzahl der Arbeitsplatzlizenzen für ein bestimmtes Produkt sowie dessen Status (aktiv/inaktiv)

2. Erstellung der Kunden-Site und Aktivierung der richtigen Produktauswahl für den Kunden. Das Konzept einer Site ist ein Container für mehrere Produkte , die für einen bestimmten Kunden lizenziert sind (z. B. Confluence und Jira Software für <site-name>.atlassian.net). Dies (siehe *Abbildung 4* unten) ist ein wichtiger Punkt, den Sie im Zusammenhang mit diesem Bericht verstehen sollten, da der Site-Container das ist, was bei diesem Vorfall gelöscht wurde, und das Konzept einer Site in diesem Dokument durchgehend diskutiert wird.

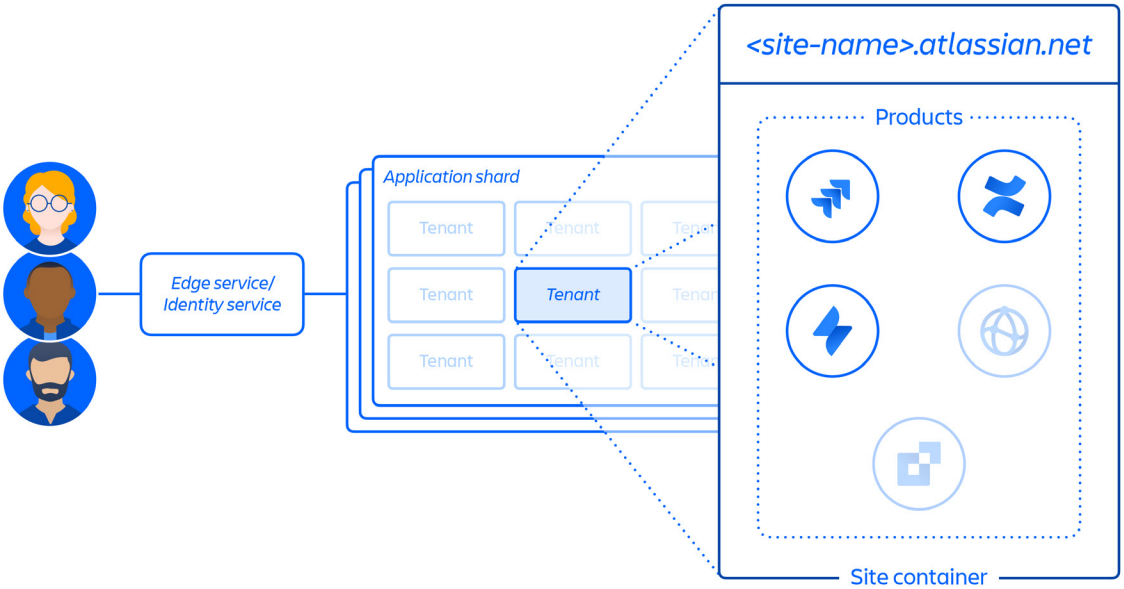


Abbildung 4: Überblick über den Site-Container

3. Bereitstellung von Produkten auf einer Kunden-Site in der angegebenen Region.

   Wenn ein Produkt bereitgestellt wird, wird der Großteil seines Inhalts in der Nähe des Zugriffsorts der Benutzer gehostet. Um die Produktleistung zu optimieren, schränken wir die Datenbewegung nicht ein, wenn Content weltweit gehostet wird. Wir können Daten nach Bedarf zwischen Regionen verschieben.

   Für einige unserer Produkte bieten wir auch Datenresidenz an. Datenresidenz ermöglicht es Kunden, zu wählen, ob Produktdaten weltweit verteilt oder an einem unserer definierten geografischen Standorte gespeichert werden.

4. Erstellung und Speicherung der Kunden-Site und der Kernmetadaten und Konfiguration der Produkte.

5 Erstellung und Speicherung der Site und Produktidentitätsdaten wie Benutzer, Gruppen, Berechtigungen usw.

6 Bereitstellung von Produktdatenbanken innerhalb einer Site, z. B. Jira-Produktfamilie, Confluence, Compass, Atlas.

7 Bereitstellung von lizenzierten Apps für das Produkt/die Produkte.

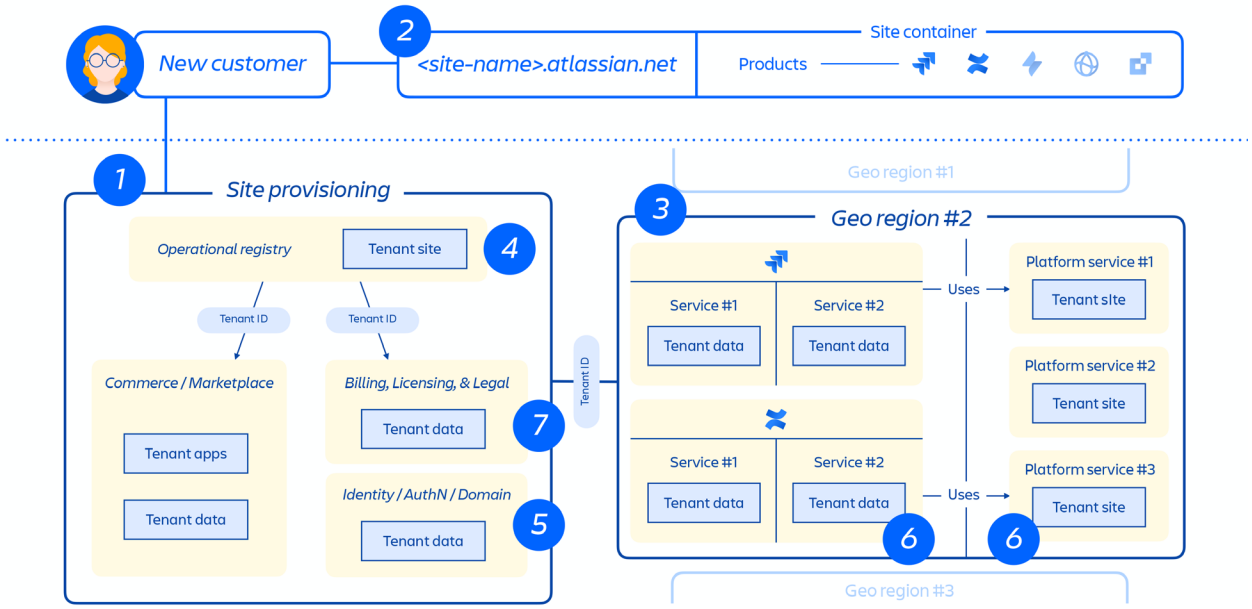


Abbildung 5: Überblick über die Bereitstellung einer Kunden-Site in einer verteilten Architektur

Abbildung 5 oben zeigt, wie die Site eines Kunden in unserer verteilten Architektur bereitgestellt wird, nicht nur in einer einzigen Datenbank oder einem Speicher. Dazu gehören mehrere physische und logische Speicherorte, an denen Metadaten, Konfigurationsdaten, Produktdaten, Plattformdaten und andere zugehörige Site-Informationen gespeichert werden.

Disaster-Recovery-Programm

Unser Disaster-Recovery-Programm (DR) umfasst alle unsere Bemühungen, Ausfallsicherheit bei Infrastrukturausfällen und Wiederherstellbarkeit von Servicespeichern aus Backups zu gewährleisten. Zwei wichtige Konzepte zum Verständnis von Disaster-Recovery-Programmen:

- **Wiederherstellungszeitziel (RTO):** Wie schnell können die Daten wiederhergestellt und im Notfall an einen Kunden zurückgegeben werden?

- **Wiederherstellungspunktziel (RPO):** Wie aktuell sind die wiederhergestellten Daten, nachdem sie aus einem Backup wiederhergestellt wurden? Wie viele Daten gehen seit der letzten Sicherung verloren?

Bei diesem Vorfall haben wir unser RTO verpasst, aber unser RPO erreicht.

Stabilität

Wir bereiten uns auf Ausfälle auf Infrastrukturebene vor, zum Beispiel den Verlust einer gesamten Datenbank, eines Service oder von AWS Availability Zones. Diese Vorbereitung umfasst die Replikation von Daten und Services über mehrere Verfügbarkeitszonen hinweg sowie regelmäßige Failover-Tests.

Wiederherstellbarkeit von Service-Speicher

Wir bereiten uns auch darauf vor, Datenbeschädigungen im Servicespeicher aufgrund von Risiken wie Ransomware, Angreifern, Softwarefehlern und Betriebsfehlern zu beheben. Diese Vorbereitung umfasst unveränderliche Backups und Tests zur Wiederherstellung von Servicespeicher-Backups. Wir sind in der Lage, jeden einzelnen Datenspeicher auf einen früheren Zeitpunkt zurückzusetzen.

Automatisierte Wiederherstellbarkeit mehrerer Sites und Produkte

Zum Zeitpunkt des Vorfalls waren wir nicht in der Lage, eine große Anzahl von Kunden-Sites auszuwählen und alle miteinander verbundenen Produkte über Backups zu einem früheren Wiederherstellungspunkt wiederherzustellen.

Unsere Möglichkeiten beschränkten sich auf Infrastruktur, Datenbeschädigung, einzelne Serviceereignisse oder Löschungen einzelner Sites. In der Vergangenheit mussten wir uns mit solchen Fehlern auseinandersetzen und Tests dazu durchführen. Für die Löschung auf Site-Ebene standen keine Runbooks zur Verfügung, die für das Ausmaß dieses Ereignisses schnell hätten automatisiert werden können. Es wären Tools und Automatisierung für alle Produkte und Services erforderlich gewesen sowie deren koordinierter Einsatz.

In den folgenden Abschnitten gehen wir näher auf diese Komplexität ein. Wir befassen uns außerdem damit, was wir bei Atlassian dafür tun, um unsere Möglichkeiten weiterzuentwickeln und zu optimieren, um diese Architektur in großem Umfang zu verwalten.

Vorfall, Zeitleiste und Wiederherstellung

Was passiert ist

2021 haben wir die Integration einer eigenständigen Atlassian-App für Jira Service Management und Jira Software namens "Insight – Asset Management" abgeschlossen. Die Funktionalität dieser eigenständigen App war dann nativ in Jira Service Management und für Jira Software nicht mehr verfügbar. Deshalb mussten wir die eigenständige Legacy-App auf Kunden-Sites löschen, auf denen sie installiert war. Unsere Entwicklungsteams verwendeten ein vorhandenes Skript und einen Prozess, um Instanzen dieser eigenständigen Anwendung zu löschen.

Es ergaben sich jedoch zwei kritische Probleme:

- **Kommunikationsfehler.** Die Kommunikation zwischen dem Team, das die Löschung veranlasst hatte, und dem Team, das mit der Durchführung betraut war, war unzureichend. Statt die IDs der zu löschenden Apps anzugeben, stellte das Team die IDs der gesamten Cloud-Site bereit, auf der die Apps gelöscht werden sollten.
- **Ungenügende Systemwarnungen.** Die API, die zum Löschen verwendet wurde, akzeptiert sowohl Site- als auch App-Kennungen und geht davon aus, dass die Eingabe korrekt ist. Dies bedeutet, dass bei Übergabe einer Site-ID eine Site gelöscht wird. Wenn eine App-ID übergeben wurde, wurde eine App gelöscht. Es gab kein Warnsignal, um die Art der angeforderten Löschung (Site oder App) zu bestätigen.

Das ausgeführte Skript folgte unserem standardmäßigen Peer-Review-Prozess, der sich darauf konzentrierte, welcher Endpunkt wie aufgerufen wurde. Die bereitgestellten Cloud-Site-IDs wurden nicht gegengeprüft, um festzustellen, ob sie sich auf die App oder auf die gesamte Site bezogen. Das Skript wurde im Staging gemäß unseren standardmäßigen Änderungsmanagementprozessen getestet. Es hätte jedoch nicht feststellen können, dass die eingegebenen IDs falsch waren, da die IDs in der Staging-Umgebung nicht existierten.

Bei der Ausführung in der Produktion lief das Skript zunächst für 30 Sites. Der erste Produktionslauf war erfolgreich und die Insight App wurde für diese 30 Sites ohne weitere Probleme gelöscht. Die IDs für diese 30 Sites wurden jedoch vor der fehlerhaften Kommunikation bezogen und enthielten die korrekten Insight-App-IDs.

Das Skript für den nachfolgenden Produktionslauf enthielt Site-IDs anstelle von Insight-App-IDs und wurde für eine Gruppe von 883 Sites ausgeführt. Die Ausführung des Skripts wurde am 5. April um 7:38 Uhr (UTC) gestartet und um 8:01 Uhr (UTC) abgeschlossen. Das

Skript löschte Sites sequenziell basierend auf der Eingabeliste, sodass die erste Kunden-Site kurz nach Beginn der Ausführung des Skripts um 7:38 Uhr (UTC) gelöscht wurde. Das Resultat war eine sofortige Löschung der 883 Sites ohne Warnung unserer Entwicklungsteams.

Die folgenden Atlassian-Produkte waren für betroffene Kunden nicht verfügbar: Jira-Produktfamilie, Confluence, Atlassian Access, Opsgenie und Statuspage.

Sobald wir von dem Vorfall erfuhren, konzentrierten sich unsere Teams auf die Wiederherstellung aller betroffenen Kunden. Zu diesem Zeitpunkt schätzten wir die Anzahl der betroffenen Sites auf etwa 700 (883 Sites waren betroffen, aber wir zogen die Sites in Besitz von Atlassian ab). Bei den 700 Sites ging es zu einem erheblichen Anteil um inaktive, kostenlose oder kleine Konten mit einer geringen Anzahl aktiver Benutzer. Basierend auf dieser Erkenntnis schätzten wir die ungefähre Anzahl der betroffenen Kunden zunächst auf rund 400.

Inzwischen haben wir viel genauere Zahlen vorliegen, weshalb Atlassian im Zuge der vollständigen Transparenz für seine Kunden bekannt gibt, dass 775 Kunden von dem Ausfall betroffen waren. Die Mehrheit der Nutzer war jedoch bereits bei der ursprünglichen Schätzung von 400 betroffenen Kunden enthalten. Für einige dieser Kunden konnte der Ausfall erst nach 14 Tagen behoben werden. Die Wiederherstellung erfolgte für die erste Kundengruppe bis zum 8. April und für alle Kunden bis zum 18. April.

Koordination

Das erste Support-Ticket von einem betroffenen Kunden wurde am 5. April um 7:46 Uhr (UTC) erstellt. Unsere interne Überwachung stellte kein Problem fest, da die Sites über einen Standard-Workflow gelöscht wurden. Um 8:17 Uhr (UTC) lösten wir unseren Vorfallmanagementprozess für schwerwiegende Vorfälle aus und beriefen ein funktionsübergreifendes Vorfallmanagementteam ein. Um 8:24 Uhr (UTC), innerhalb von sieben Minuten, wurde der Vorfall als "kritisch" eingestuft. Um 8:53 Uhr (UTC) bestätigte unser Team, dass das Kundensupport-Ticket und die Skriptausführung zusammenhängen. Nachdem wir die Komplexität der Wiederherstellung erkannt hatten, wiesen wir dem Vorfall um 12:38 Uhr (UTC) den höchsten Schweregrad zu.

Das Vorfallmanagementteam bestand aus Mitarbeitern mehrerer Teams von Atlassian, darunter Engineering, Kundensupport, Programmmanagement, Kommunikation und viele mehr. Solange der Vorfall nicht gelöst war, traf sich das Kernteam alle drei Stunden, bis alle Sites wiederhergestellt und validiert waren und den Kunden übergeben werden konnten.

Um den Wiederherstellungsfortschritt zu verwalten, erstellten wir ein neues Jira-Projekt, SITE und einen Workflow zur Nachverfolgung der Wiederherstellung für jede einzelne Site über mehrere Teams hinweg (Engineering, Programmmanagement, Support usw.). Dieser Ansatz ermöglichte es allen Teams, Probleme im Zusammenhang mit der Wiederherstellung einzelner Sites auf einfache Weise zu identifizieren und zu verfolgen.

Außerdem haben wir für die Dauer des Vorfalls am 8. April um 3:30 Uhr (UTC) einen Code-Freeze für die gesamte Technik implementiert. So konnten wir uns auf die Wiederherstellung von Kunden konzentrieren, das Risiko von Änderungen und damit einhergehenden Inkonsistenzen in Kundendaten eliminieren, das Risiko anderer Ausfälle minimieren und die Wahrscheinlichkeit verringern, dass sonstige Änderungen das Team von der Wiederherstellung ablenken.

Zeitleiste des Vorfalls

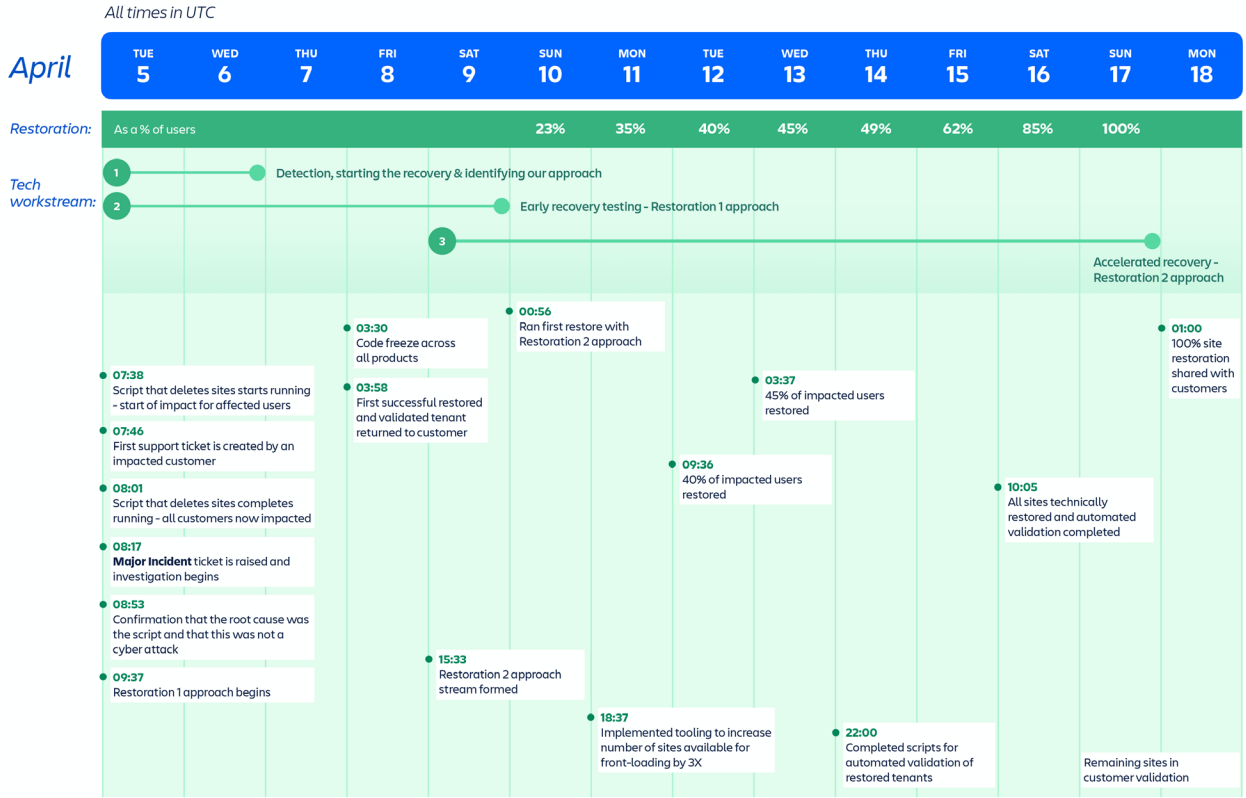


Abbildung 6: Zeitleiste des Vorfalls und wichtige Meilensteine der Wiederherstellung

Allgemeiner Überblick über Wiederherstellungs-Workstreams

Die Wiederherstellung lief in drei primären Workstreams ab: Erkennung, frühe Wiederherstellung und Beschleunigung. Obwohl wir im Folgenden jeden Workstream separat beschreiben, wurde während der Wiederherstellung die Arbeit in allen Workstreams parallel ausgeführt.

Workstream 1: Erkennung, Start der Wiederherstellung und Identifizierung unseres Ansatzes

Zeitstempel: Tage 1 bis 2 (5. bis 6. April)
Um 8:53 Uhr (UTC) am 5. April stellten wir fest, dass das Insight-App-Skript das Löschen von Sites verursacht hat. Wir stellten sicher, dass dies nicht das Ergebnis einer internen böswilligen Handlung oder eines Cyberangriffs war. Relevante Teams für Produkt- und Plattforminfrastruktur wurden in die Behebung des Vorfalls einbezogen.

Zu Beginn des Vorfalls stellten wir Folgendes fest:

- Das Wiederherstellen von Hunderten gelöschter Sites ist ein komplexer, mehrstufiger Prozess (detaillierte Angaben dazu finden Sie im Abschnitt "Architektur" oben), für dessen Erfolg mehrere Teams und mehrere Tage nötig sind.
- Wir hatten die Möglichkeit, eine einzelne Site wiederherzustellen, aber wir hatten keine Funktionen und Prozesse für die Wiederherstellung einer großen Anzahl von Sites.

Daher mussten wir den Wiederherstellungsprozess erheblich parallelisieren und automatisieren, um betroffenen Kunden zu helfen, so schnell wie möglich wieder Zugang zu ihren Atlassian-Produkten zu erhalten.

Bei Workstream 1 arbeiteten mehrere Entwicklungsteams an den folgenden Aktivitäten:

- Identifizieren und Ausführen von Wiederherstellungsschritten für Site-Batches in der Pipeline.
- Schreiben und Verbessern der Automatisierung, damit die Teams Wiederherstellungsschritte für eine größere Anzahl von Sites in einem Batch ausführen können.

Workstream 2: Frühe Wiederherstellung und der Ansatz "Wiederherstellung 1"

Zeitstempel: Tage 1 bis 4 (5. bis 9. April)
Wir haben innerhalb einer Stunde, nachdem das Skript seine Ausführung beendet hat, verstanden, was die Löschung der Sites am 5. April um 8:53 Uhr (UTC) verursacht hat.
Wir identifizierten auch den Wiederherstellungsprozess, der zuvor verwendet wurde, um eine

kleine Anzahl von Sites wieder in Produktion zu bringen. Der Wiederherstellungsprozess für die Wiederherstellung gelöschter Sites in einem solchen Ausmaß war jedoch nicht gut definiert.

Um schnell voranzukommen, wurden die frühen Phasen des Vorfalls in zwei Arbeitsgruppen aufgeteilt:

- Die manuelle Arbeitsgruppe validierte die erforderlichen Schritte und führte den Wiederherstellungsprozess für eine kleine Anzahl von Sites manuell aus.
- Die Arbeitsgruppe "Automatisierung" übernahm den bestehenden Wiederherstellungsprozess und integrierte die Automatisierung, um die Schritte für größere Site-Batches auszuführen.

Überblick über den Ansatz "Wiederherstellung 1" (siehe *Abbildung 7* unten):

- Es erforderte die Erstellung einer neuen Site für jede gelöschte Site, gefolgt von jedem nachgelagerten Produkt, Service und Datenspeicher, für die Daten wiederhergestellt werden mussten.
- Die neue Site würde neue IDs wie `CloudID` enthalten. Diese IDs sind unveränderlich, was bedeutet, dass viele Systeme diese IDs in Datensätze einbetten. Wenn sich diese IDs ändern, müssen wir große Datenmengen aktualisieren, was besonders bei Ökosystem-Apps von Drittanbietern problematisch ist.
- Beim Ändern einer neuen Site, um den Status der gelöschten Site zu replizieren, traten komplexe und oft unvorhersehbare Abhängigkeiten zwischen den Schritten auf.

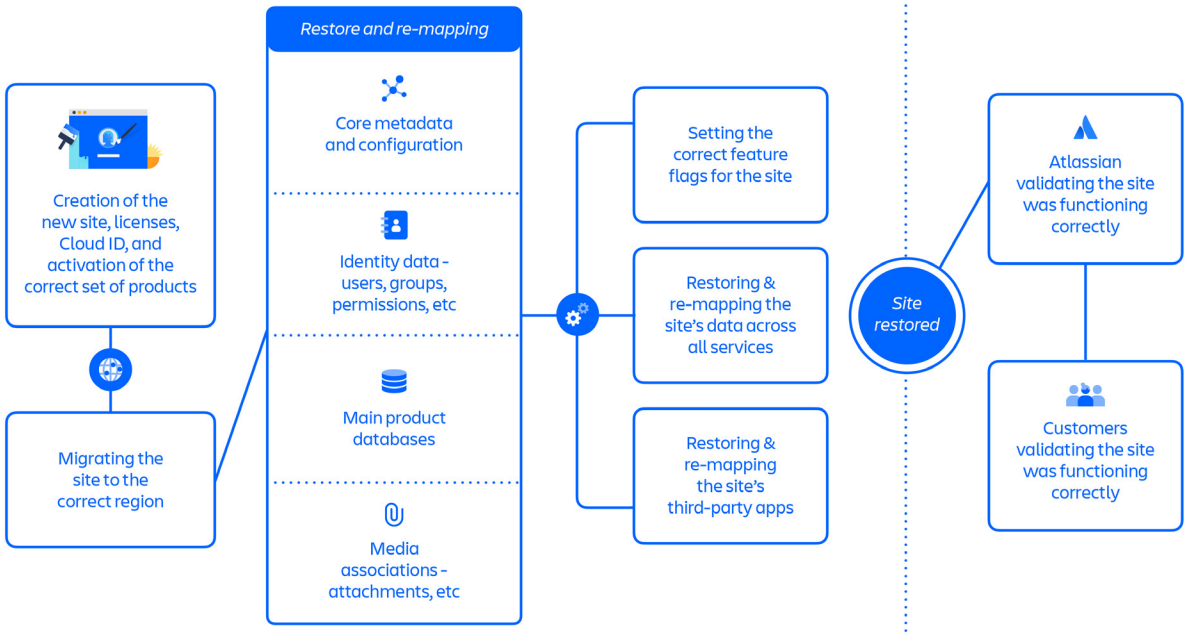


Abbildung 7: Die wichtigsten Schritte im Ansatz "Wiederherstellung 1".

Der Ansatz "Wiederherstellung 1" umfasste ungefähr 70 einzelne Schritte. Fasst man diese auf hoher Ebene zusammen, führt dies zu folgendem Ablauf:

- Erstellung der neuen Site, Lizenzen, Cloud-ID und Aktivierung der richtigen Produkte
- Migration der Site in die richtige Region
- Wiederherstellung und Neuzuordnung der Kernmetadaten und Konfiguration der Site
- Wiederherstellung und Neuzuordnung der Identitätsdaten der Site – Benutzer, Gruppen, Berechtigungen usw.
- Wiederherstellung der wichtigsten Produktdatenbanken der Site
- Wiederherstellung und Neuzuordnung der Medienzuordnungen der Site – Anhänge usw.
- Einstellung der richtigen Feature-Flags für die Site
- Wiederherstellung und Neuzuordnung der Site-Daten über alle Services hinweg
- Wiederherstellung und Neuzuordnung der Drittanbieter-Apps der Site
- Überprüfung durch Atlassian, ob die Site ordnungsgemäß funktioniert
- Überprüfung durch die Kunden, ob die Site ordnungsgemäß funktioniert

Nach der Optimierung dauerte der Ansatz "Wiederherstellung 1" etwa 48 Stunden, um einen Batch von Sites wiederherzustellen. Dieser Ansatz wurde zwischen dem 5. April und dem 14. April für die Wiederherstellung von 53 % der betroffenen Benutzer für 112 Sites verwendet.

Workstream 3: Beschleunigte Wiederherstellung und der Ansatz "Wiederherstellung 2"

Zeitstempel: Tage 4 bis 13 (9. bis 17. April)
Mit dem Ansatz "Wiederherstellung 1" hätten wir drei Wochen gebraucht, um alle Kunden wiederherzustellen. Daher haben wir am 9. April einen neuen Ansatz vorgeschlagen, um die Wiederherstellung aller Sites zu beschleunigen – Wiederherstellung 2 (siehe *Abbildung 8* unten).

Der Ansatz "Wiederherstellung 2" bot eine verbesserte Parallelität zwischen Wiederherstellungsschritten, indem die Komplexität und die Anzahl der Abhängigkeiten reduziert wurden, die beim Ansatz "Wiederherstellung 1" vorhanden waren.

"Wiederherstellung 2" beinhaltete die Neuerstellung (oder Aufhebung der Löschung) von Datensätzen, die mit der Site verknüpft sind, über alle entsprechenden Systeme hinweg, beginnend mit dem Catalogue-Service-Datensatz. Ein Schlüsselelement dieses neuen Ansatzes war die *Wiederverwendung aller Site-IDs der alten Site*. Dadurch wurde mehr als die Hälfte der Schritte aus dem vorherigen Prozess entfernt, die verwendet wurden, um die alten IDs den neuen IDs zuzuordnen , einschließlich der Notwendigkeit, sich mit jedem Drittanbieter von Apps für jede Site abzustimmen.

Die Umstellung von "Wiederherstellung 1" auf "Wiederherstellung 2" führte jedoch zu einem erheblichen Mehraufwand bei der Incident Response:

- Viele der Automatisierungsskripte und -prozesse, die im Ansatz "Wiederherstellung 1" eingerichtet wurden, mussten für "Wiederherstellung 2" geändert werden.
- Teams, die Wiederherstellungen durchführten (einschließlich Vorfallkoordinatoren), mussten bei beiden Ansätzen parallele Wiederherstellungs-Batches verwalten, während wir den Ablauf von "Wiederherstellung 2" testeten und validierten.
- Die Verwendung eines neuen Ansatzes bedeutete, dass wir den Prozess "Wiederherstellung 2" testen und validieren mussten, bevor wir ihn skalieren konnten. Das bedeutete wiederum, dass die Validierung, die bereits für "Wiederherstellung 1" abgeschlossen wurde, erneut durchgeführt werden musste.

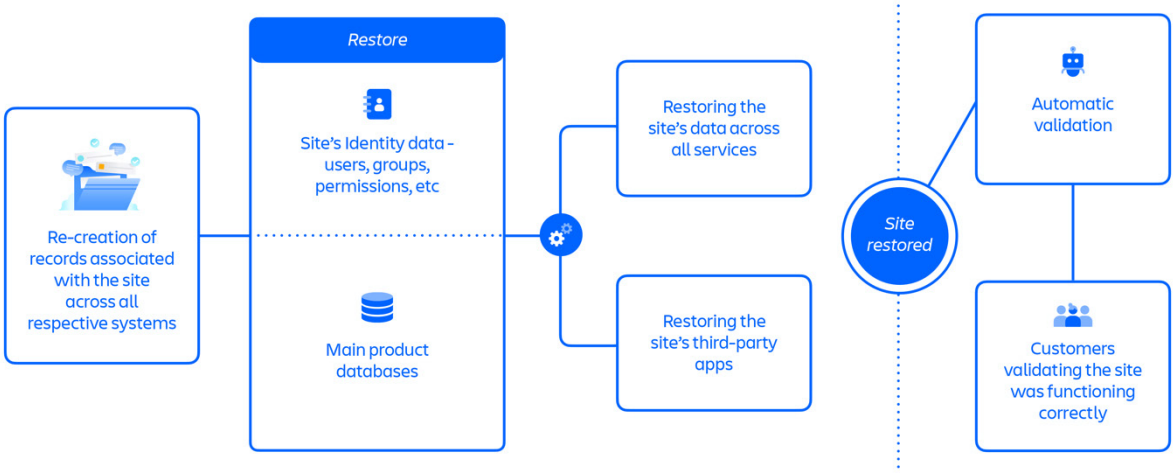


Abbildung 8: Die wichtigsten Schritte im Ansatz "Wiederherstellung 2"

Die Grafik oben stellt den Ansatz "Wiederherstellung 2" dar, der über 30 Schritte umfasste, die dem folgenden weitgehend parallelisierten Ablauf folgten:

- Wiederherstellung von Datensätzen, die mit der Site verknüpft sind, über alle entsprechenden Systeme hinweg
- Wiederherstellung der Identitätsdaten der Site – Benutzer, Gruppen, Berechtigungen usw.
- Wiederherstellung der wichtigsten Produktdatenbanken der Site
- Wiederherstellung der Daten der Site über alle Services hinweg
- Wiederherstellung der Drittanbieter-Apps der Site
- Automatische Validierung
- Überprüfung durch die Kunden, ob die Site ordnungsgemäß funktioniert

Im Rahmen der beschleunigten Wiederherstellung griffen wir auch zu Frontloading und zur Automatisierung der Site-Wiederherstellung, da die manuelle Wiederherstellung für große Batches nicht gut skaliert werden kann. Aufgrund der sequenziellen Natur des Wiederherstellungsprozesses könnte die Site-Wiederherstellung bei großen

Datenbankwiederherstellungen und Wiederherstellungen der Benutzerbasis/Berechtigungen langsamer sein. Implementierte Optimierungen:

- Wir entwickelten die Tools und Richtlinien, die für das *Frontloading* und die lange Ausführung von Schritten wie Datenbankwiederherstellungen und Identitätssynchronisationen erforderlich sind, damit sie vor anderen Wiederherstellungsschritten abgeschlossen werden konnten.
- Entwicklerteams integrierten Automatisierung in die einzelnen Schritte, damit große Wiederherstellungs-Batches sicher ausgeführt werden konnten.
- Die Automatisierung wurde integriert, um sicherzustellen, dass Sites nach Abschluss aller Wiederherstellungsschritte ordnungsgemäß funktionierten.

Bei dem beschleunigten Ansatz "Wiederherstellung 2" dauerte die Wiederherstellung einer Site etwa 12 Stunden. Dieser Ansatz wurde zwischen dem 14. und 17. April für die Wiederherstellung von etwa 771 Sites von etwa 47 % der betroffenen Benutzer verwendet.

Minimaler Datenverlust nach der Wiederherstellung gelöschter Sites

Unsere Datenbanken werden mit einer Kombination aus vollständigen Backups und inkrementellen Backups gesichert, die es uns ermöglichen, einen bestimmten Zeitpunkt für die Wiederherstellung unserer Datenspeicher innerhalb der Backup-Aufbewahrungsfrist (30 Tage) auszuwählen. Für die meisten Kunden identifizierten wir während des Vorfalls die wichtigsten Datenspeicher für unsere Produkte und entschieden uns, einen Wiederherstellungspunkt von fünf Minuten vor dem Löschen der Sites als sicheren Synchronisationspunkt zu verwenden. Die nicht primären Datenspeicher wurden an demselben Punkt oder durch Wiederholung der aufgezeichneten Ereignisse wiederhergestellt. Die Verwendung eines festen Wiederherstellungspunkts für primäre Speicher ermöglichte es uns, die Konsistenz der Daten über alle Datenspeicher hinweg zu erreichen.

Bei 57 Kunden, die zu einem frühen Zeitpunkt im Rahmen der Incident Response wiederhergestellt wurden, führte das Fehlen konsistenter Richtlinien und das manuelle Abrufen von Datenbank-Backup-Snapshots dazu, dass einige Confluence- und Insight-Datenbanken auf einen Zeitpunkt wiederhergestellt wurden, der *mehr* als fünf Minuten vor dem Zeitpunkt der Site-Löschung liegt. Diese Inkonsistenz wurde während eines Auditprozesses nach der Wiederherstellung festgestellt. Inzwischen haben wir den Rest der Daten wiederhergestellt und die davon betroffenen Kunden kontaktiert. Wir helfen ihnen dabei, Änderungen vorzunehmen, um ihre Daten vollständig wiederherzustellen.

Fazit:

- Wir haben während dieses Vorfalls unser Wiederherstellungspunktziel (RPO) von einer Stunde erreicht.

- Der Datenverlust aufgrund des Vorfalls ist auf fünf Minuten vor dem Löschen der Site begrenzt.
- Bei einer geringen Anzahl von Kunden wurden Confluence- oder Insight-Datenbanken auf einen Punkt wiederhergestellt, der mehr als fünf Minuten vor dem Löschen der Site zurückliegt. Wir können die Daten jedoch wiederherstellen und arbeiten derzeit mit den Kunden daran, diese Daten wiederherzustellen.

Kommunikation bei Vorfällen

Die Kommunikation bei Vorfällen umfasst den Kontakt mit Kunden, Partnern, den Medien, Branchenanalysten, Investoren und der allgemeinen Technologiegemeinschaft.

Was passiert ist

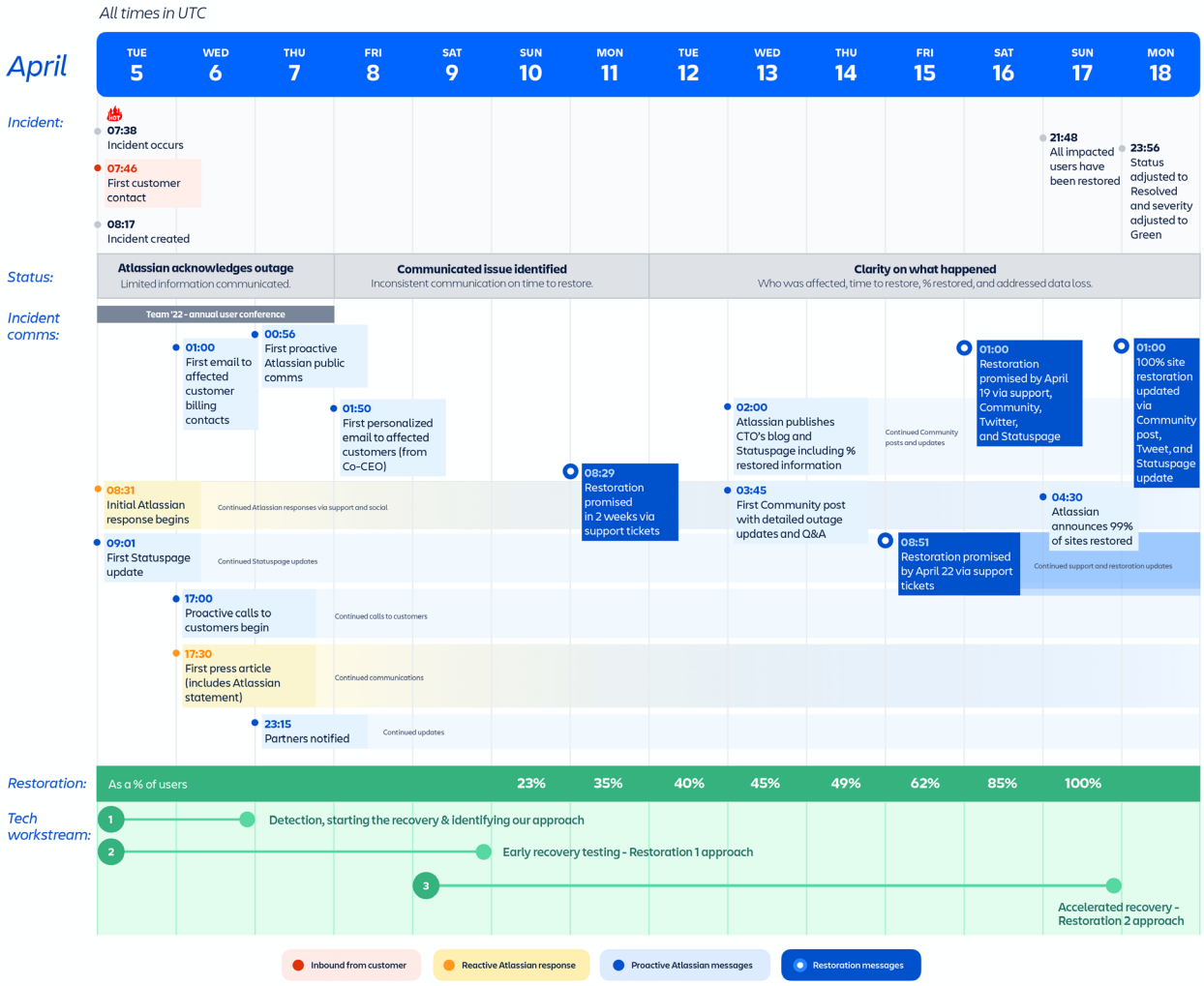


Abbildung 9: Zeitleiste der wichtigsten Meilensteine der Kommunikation bei diesem Vorfall

Zeitstempel: Tage 1 bis 3 (5. bis 7. April)

Frühe Reaktion

Das erste Support-Ticket wurde am 5. April um 7:46 Uhr (UTC) erstellt und der Atlassian-Support antwortete mit der Bestätigung des Vorfalls um 8:31 Uhr (UTC). Um 9:03 Uhr (UTC) wurde das erste Statuspage-Update veröffentlicht, um Kunden darüber zu informieren, dass wir den Vorfall untersuchen. Um 11:13 Uhr (UTC) bestätigten wir über Statuspage, dass wir die Ursache identifizieren konnten und an einer Lösung arbeiten. Um 1:00 Uhr (UTC) am 6. April wurde in den ersten Mitteilungen zum Kundenticket angegeben, dass der Ausfall auf ein Wartungsskript zurückzuführen ist und wir einen minimalen Datenverlust erwarten. Atlassian antwortete auf Medienanfragen am 6. April um 17:30 Uhr (UTC) mit einer Erklärung. Atlassian twitterte seine erste umfassende externe Nachricht, in der der Vorfall am 7. April um 00:56 Uhr (UTC) bestätigt wurde.

Zeitstempel: Tage 4 bis 7 (8. bis 11. April)

Eine umfassendere, personalisierte Öffentlichkeitsarbeit beginnt

Am 8. April um 1:50 Uhr (UTC) schickte Atlassian den betroffenen Kunden eine E-Mail mit einer Entschuldigung von Mitbegründer und Co-CEO Scott Farquhar. In den folgenden Tagen arbeiteten wir daran, die gelöschten Kontaktinformationen wiederherzustellen und Support-Tickets für alle betroffenen Sites zu erstellen, für die noch kein Support-Ticket eingereicht wurde. Unser Support-Team sendete dann weiterhin regelmäßig Updates über die Wiederherstellungsmaßnahmen über die Support-Tickets zu den einzelnen betroffenen Sites.

Zeitstempel: Tage 8 bis 14 (12. bis 18. April)

Größere Klarheit und vollständige Wiederherstellung

Am 12. April veröffentlichte Atlassian ein Update von CTO, Sri Viswanath, der weitere technische Details zu dem Vorfall bekannt gab sowie wer betroffen war, ob es zu Datenverlusten gekommen war, welche Fortschritte wir bei der Wiederherstellung machten und dass es möglicherweise zwei Wochen dauern würde, um alle Sites vollständig wiederherzustellen. Der Blog wurde von einer weiteren Presseerklärung von Sri begleitet. In unserem ersten proaktiven Beitrag der Atlassian Community von Stephen Deasy, Head of Engineering, verwiesen wir ebenso auf Sris Blog. Infolgedessen wurde der Blog zu der Kommunikationsplattform, über die Updates bekanntgegeben sowie Fragen und Antworten der breiten Öffentlichkeit beantwortet wurden. Am 18. April konnten wir über diesen Beitrag die vollständige Wiederherstellung aller betroffenen Kunden-Sites bekanntgeben.

Warum haben wir nicht früher öffentlich reagiert?

1. Wir haben die direkte Kommunikation mit betroffenen Kunden über Statuspage, E-Mail, Support-Tickets und direkte Gespräche priorisiert. Wir konnten jedoch viele Kunden nicht erreichen, da wir ihre Kontaktinformationen verloren hatten, als ihre Sites gelöscht wurden. Wir hätten viel früher eine umfassendere Kommunikation implementieren müssen, um die betroffenen Kunden und Endnutzer über unserer Incident Response und die Zeitleiste für die Problembehebung zu informieren.

2. Wir wussten zwar sofort, was den Vorfall verursacht hatte, doch die komplexe Architektur und die einzigartigen Umstände dieses Vorfalls machten es uns schwer, den Zeitraum bis zur Fehlerbehebung einzuschätzen. Statt abzuwarten, bis wir einen vollständigen Überblick hatten, hätten wir transparent darüber informieren müssen, was wir wussten und was wir nicht wussten. Die Bereitstellung einer allgemeinen Schätzung der Wiederherstellungsdauer (selbst wenn diese nur ungefähr gewesen wäre) sowie eine klare Aussage darüber, wann wir in etwa einen klaren Überblick über den Vorfall haben werden, hätte unseren Kunden eine bessere Planung in Bezug auf den Vorfall ermöglicht. Dies gilt insbesondere für Systemadministratoren und technische Ansprechpartner, die bei der Verwaltung von Stakeholdern und Benutzern in ihren Organisationen an vorderster Front stehen.

Support und Kundenkontakt

Wie bereits erwähnt, löschte dasselbe Skript, das Kunden-Sites gelöscht hatte auch wichtige Kunden-IDs und Kontaktinformationen (z. B. Cloud-URL, Site-Systemadministrator-Kontakte) aus unseren Produktionsumgebungen. Dies ist bemerkenswert, weil unsere Kernsysteme (z. B. Support, Lizenzierung, Abrechnung) alle die Existenz einer Cloud-URL und Site-Systemadministrator-Kontakte als primäre Identifikatoren für Sicherheits-, Routing- und Priorisierungszwecke nutzen. Als wir diese Identifikatoren verloren, verloren wir zunächst unsere Fähigkeit, Kunden systematisch zu identifizieren und mit ihnen in Kontakt zu treten.

Inwiefern war der Support für unsere Kunden beeinträchtigt?

Erstens: Die Mehrheit der betroffenen Kunden konnte unser Support-Team nicht über das normale Online-Kontaktformular erreichen. Dieses Formular ist so konzipiert, dass sich ein Benutzer mit seiner Atlassian-ID anmelden und eine gültige Cloud-URL angeben muss. Ohne eine gültige URL kann der Benutzer kein Ticket für den technischen Support einreichen. Im normalen Geschäftsverlauf ist diese Überprüfung für die Sicherheit der Site und die Ticketsuche gedacht. Diese Anforderung führte jedoch zu einem unbeabsichtigten Problem für Kunden, die von diesem Ausfall betroffen waren. Sie wurden dadurch daran gehindert, ein Site-Support-Ticket mit hoher Priorität einzureichen.

Zweitens: Die durch den Vorfall verursachte Löschung der Systemadministrator-Daten der Site verhinderte eine proaktive Kontaktaufnahme mit betroffenen Kunden. In den ersten Tagen des Vorfalls sendeten wir proaktiv Mitteilungen an die bei Atlassian registrierten Abrechnungs- und technischen Ansprechpartner der betroffenen Kunden. Wir stellten jedoch schnell fest, dass viele Abrechnungs- und technische Kontakte für die betroffenen Kunden veraltet waren. Ohne die Informationen zum Systemadministrator für jede Site verfügten wir nicht über eine vollständige Liste der aktiven und genehmigten Kontakte für die Kontaktaufnahme.

Wie haben wir reagiert?

Unsere Support-Teams konzentrierten sich in den ersten Tagen des Vorfalls auf drei gleichermaßen wichtige Prioritäten, um die Wiederherstellung der Sites zu beschleunigen und unsere Kommunikationskanäle wiederherzustellen.

Erstens: Erstellen einer zuverlässigen Liste mit bestätigten Kundenkontakten. Während unsere Entwicklungsteams an der Wiederherstellung der Sites arbeiteten, konzentrierten sich unsere kundenorientierten Teams auf die Wiederherstellung bestätigter Kontaktinformationen. Wir nutzten alle uns zur Verfügung stehenden Mittel (Abrechnungssysteme, alte Support-Tickets, andere gesicherte Benutzer-Backups, direkter Kundenkontakt usw.), um unsere Liste mit Kontakten neu zu erstellen. Unser Ziel war es, für jede betroffene Site ein Support-Ticket im Zusammenhang mit dem Vorfall zu haben, um die direkte Kontaktaufnahme und Reaktionszeiten zu optimieren.

Zweitens: Wiederherstellung von für diesen Vorfall spezifischen Workflows, Warteschlangen und SLAs. Das Löschen der Cloud-ID und die Unfähigkeit, Benutzer korrekt zu authentifizieren, wirkte sich auch auf unsere Fähigkeit aus, Support-Tickets im Zusammenhang mit dem Vorfall über unsere normalen Systeme zu verarbeiten. Tickets wurden nicht mit der relevanten Priorität, in den korrekten Warteschlangen und Dashboards angezeigt. Wir stellten schnell ein funktionsübergreifendes Team (Support, Produkt, IT) zusammen, um zusätzliche Logik, SLAs, Workflow-Status und Dashboards zu entwerfen und

hinzuzufügen. Da dies innerhalb unseres Produktionssystems erfolgen musste, dauerte die vollständige Entwicklung, das Testen und die Bereitstellung mehrere Tage.

Drittens: Umfassende Skalierung manueller Validierungen zur Beschleunigung der Site-Wiederherstellung. Als die Techniker bei den ersten Wiederherstellungen Fortschritte machten, wurde klar, dass die Kapazität unserer globalen Support-Teams erforderlich sein würde, um die Site-Wiederherstellung durch manuelle Tests und Validierungsprüfungen zu beschleunigen. Dieser Validierungsprozess würde zu einem wichtigen Weg werden, um unseren Kunden wiederhergestellte Sites zur Verfügung zu stellen, sobald unser Entwicklungsteam die Datenwiederherstellung beschleunigt. Wir mussten unabhängige Standardarbeitsanweisungen (SOPs), Workflows, Übergaben und Personalpläne erstellen, um mehr als 450 Support-Techniker für die Durchführung von Validierungsprüfungen zu mobilisieren. Durch Schichtarbeit wurde eine Verfügbarkeit rund um die Uhr sichergestellt, um die Wiederherstellung für unsere Kunden zu beschleunigen.

Obwohl diese wichtigen Prioritäten bis zum Ende der ersten Woche festgelegt waren, konnten wir kaum *aussagekräftige* Updates geben, da aufgrund der Komplexität der Wiederherstellung kein genauer Zeitrahmen für die Behebung des Vorfalls eingegrenzt werden konnte. Wir hätten früher zu unserer Unsicherheit bei der Veröffentlichung eines Wiederherstellungsdatums stehen und früher persönliche Gespräche suchen sollen, damit unsere Kunden entsprechend hätten planen können.

Wie werden wir uns verbessern?

Wir haben Massenlöschungen von Sites sofort blockiert, bis entsprechende Änderungen vorgenommen werden können.

Nach der Bewältigung dieses Vorfalls und der Neubewertung unserer internen Prozesse möchte ich feststellen, dass es nicht die Menschen sind, die Vorfälle verursachen. Es sind Lücken in den Systemen, die Fehler erst möglich machen. In diesem Abschnitt werden die Faktoren zusammengefasst, die zu diesem Vorfall beigetragen haben. Wir besprechen auch unsere Pläne für eine schnellere Lösung dieser Schwächen und Probleme.

Erkenntnis 1: "Vorläufiges Löschen" sollte bei allen Systemen eingesetzt werden

Eine Löschung in diesem Umfang sollte verboten werden oder mehrere Sicherheitsebenen haben, um Fehler zu vermeiden. Die wichtigste Verbesserung, die wir vornehmen werden, besteht darin, die Löschung von Kundendaten und Metadaten zu verhindern, die zuvor noch nicht vorläufig gelöscht wurden.

a) Das Löschen von Daten sollte nur in Form einer vorläufigen Löschung erfolgen
Das Löschen einer gesamten Site sollte verboten werden. Bei vorläufigen Löschvorgängen sollten Schutzmaßnahmen auf mehreren Ebenen integriert werden, um Fehler zu vermeiden. Wir werden eine Richtlinie für vorläufiges Löschen implementieren, die verhindert, dass externe Skripts oder Systeme Kundendaten in einer Produktionsumgebung löschen. Die Richtlinie für vorläufiges Löschen wird eine ausreichende Datenaufbewahrung ermöglichen, damit Daten schnell und sicher wiederhergestellt werden können. Die Daten werden erst nach Ablauf einer Aufbewahrungsfrist aus der Produktionsumgebung gelöscht.

Maßnahmen:

- **Implementieren vorläufiger Löschungen in den Bereitstellungs-Workflows sowie allen relevanten Datenspeichern:** Darüber hinaus stellt das Tenant-Platform-Team sicher, dass Datenlöschungen nur nach Deaktivierungen sowie nach anderen Sicherheitsvorkehrungen in diesem Bereich erfolgen können. Langfristig wird die Tenant Platform eine führende Rolle bei der Weiterentwicklung des korrekten Statusmanagements von Mandantendaten übernehmen.

b) Es sollte ein standardisierter und verifizierter Überprüfungsprozess für vorübergehende Löschvorgänge vorhanden sein
Vorübergehende Löschvorgänge bergen ein hohes Risiko. Daher sollten standardisierte oder automatisierte Überprüfungsprozesse vorhanden sein, die definierte Rollbacks und Testverfahren beinhalten, um diese Vorgänge anzuleiten.

Maßnahmen:

- **Erzwungener schrittweiser Rollout aller vorübergehender Löschungen:** Alle neuen Vorgänge, die gelöscht werden müssen, werden zunächst auf unseren eigenen Sites getestet, um unseren Ansatz zu validieren und die Automatisierung zu überprüfen. Sobald wir diese Validierung abgeschlossen haben, werden wir die Kunden schrittweise durch denselben Prozess führen und weiterhin auf Unregelmäßigkeiten testen, bevor wir die Automatisierung auf die gesamte ausgewählte Benutzerbasis anwenden.

- **Für vorläufige Löschungen muss ein getesteter Rollback-Plan bereitstehen:** Für jede Aktivität in Zusammenhang mit der vorläufigen Löschung von Daten muss zuerst die Wiederherstellung der gelöschten Daten getestet werden, bevor die Aktion in der Produktion ausgeführt werden kann. Außerdem muss ein getesteter Rollback-Plan vorhanden sein.

Erkenntnis 2: Im Rahmen des DR-Programms sollte die Wiederherstellung bei Löschvorgängen mehrerer Sites und Produkte bei diversen Kunden automatisiert werden

Atlassian Data Management beschreibt unseren Datenverwaltungsprozess im Detail. Um eine hohe Verfügbarkeit sicherzustellen, stellen wir ein synchrones Standby-Replikat in mehreren AWS Availability Zones (AZ) bereit und warten dieses. Das AZ-Failover ist automatisiert und dauert in der Regel 60 bis 120 Sekunden. Wir beheben regelmäßig Rechenzentrumsausfälle und andere häufige Störungen, ohne dass diese Auswirkungen auf Kunden haben.

Zudem verfügen wir über unveränderliche Backups, die vor Datenbeschädigungen geschützt sind und eine Wiederherstellung auf einen früheren Zeitpunkt ermöglichen. Backups werden 30 Tage lang gespeichert und Atlassian prüft sie regelmäßig auf ihre Wiederherstellungsfähigkeit. Bei Bedarf können wir alle Kunden in einer neuen Umgebung wiederherstellen.

Mit diesen Backups stellen wir regelmäßig die Daten einzelner Kunden oder Kundengruppen wieder her, die versehentlich ihre Daten gelöscht haben. Für die Löschung auf Site-Ebene standen jedoch keine Runbooks zur Verfügung, die für das Ausmaß dieses Ereignisses schnell hätten automatisiert werden können. Es wären Tools und Automatisierung für alle Produkte und Services erforderlich gewesen sowie deren koordinierter Einsatz.

Was wir (noch) nicht automatisiert haben, ist die Wiederherstellung einer großen Teilmenge von Kunden in unsere bestehende (und derzeit verwendete) Umgebung, ohne einen unserer anderen Kunden zu beeinträchtigen.

In unserer Cloud-Umgebung enthält jeder Datenspeicher Daten von mehreren Kunden. Da die bei diesem Vorfall gelöschten Daten nur einen Teil der Datenspeicher ausmachten, die weiterhin von anderen Kunden verwendet werden, müssen wir einzelne Daten aus unseren Backups manuell extrahieren und wiederherstellen. Jede Wiederherstellung einer Kunden-Site ist ein langwieriger und komplexer Prozess, der eine interne Überprüfung und eine abschließende Kundenverifizierung erfordert, nachdem die Site wiederhergestellt wurde.

Maßnahmen:

- **Beschleunigung der Wiederherstellung mehrerer Produkte und Sites für eine größere Gruppe von Kunden:** Das DR-Programm erfüllt unsere aktuellen RPO-Standards von einer Stunde. Wir werden die Automatisierung und die Erkenntnisse aus diesem Vorfall nutzen, um das DR-Programm zu beschleunigen und das RTO gemäß unserer Richtlinie für Vorfälle diese Größenordnung zu erfüllen.

Automatisieren der Überprüfung dieses Falls und Hinzufügen zu DR-Tests: Wir führen regelmäßig DR-Übungen durch, bei denen alle Produkte für mehrere Sites wiederhergestellt werden. Diese DR-Tests werden überprüfen, ob Runbooks auf dem neuesten Stand sind, wenn sich unsere Architektur weiterentwickelt und neue Grenzfälle auftreten. Wir werden unseren Wiederherstellungsansatz kontinuierlich verbessern, den Wiederherstellungsprozess weiter automatisieren und die Wiederherstellungszeit reduzieren.

Erkenntnis 3: Verbesserung des Vorfallmanagements bei schwerwiegenden Vorfällen

Unser Vorfallmanagementprogramm eignet sich hervorragend für die Bewältigung der größeren und kleineren Vorfälle, die im Laufe der Jahre aufgetreten sind. Wir simulieren häufig eine Incident Response für Vorfälle kleinerer Größenordnung mit kürzerer Dauer, an denen in der Regel weniger Personen und Teams beteiligt sind.

Zu Spitzenzeiten arbeiteten bei diesem Vorfall jedoch Hunderte von Technikern und Mitarbeitern des Kundensupports gleichzeitig an der Wiederherstellung von Kunden-Sites. Unser Vorfallmanagementprogramm und unsere Teams waren nicht auf die Größe, den Umfang und die Dauer dieser Art von Vorfällen vorbereitet (siehe *Abbildung 10* unten).

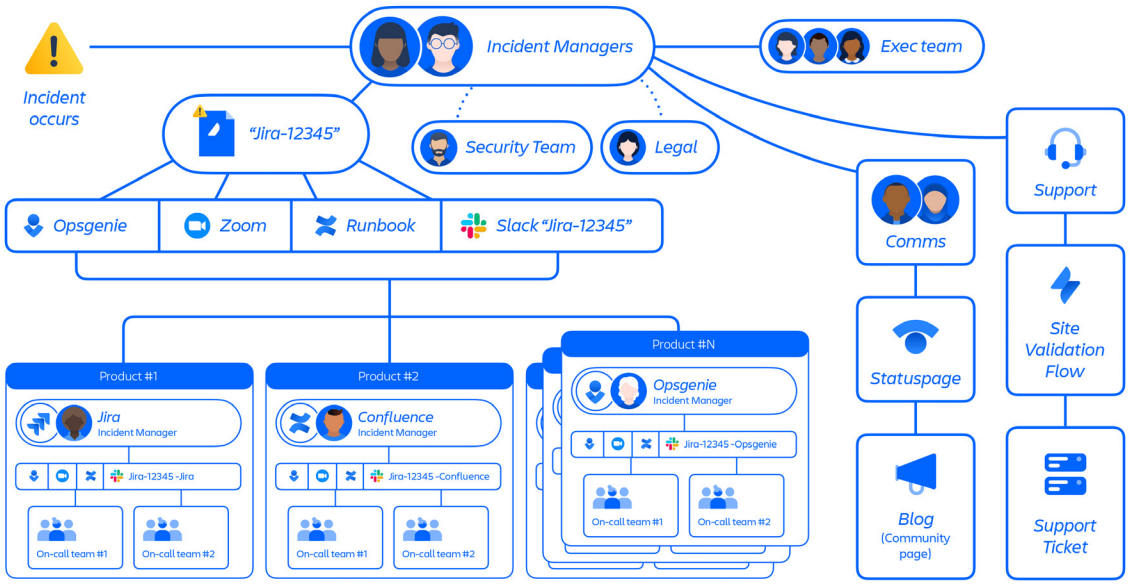


Abbildung 10: Überblick über den umfangreichen Vorfallmanagementprozess

Unser groß angelegter Vorfallmanagementprozess wird besser definiert und häufig geübt

Wir haben Playbooks für Vorfälle auf Produktebene, aber nicht für Ereignisse dieser Größenordnung, bei denen Hunderte von Personen gleichzeitig im gesamten Unternehmen arbeiten. Zu unseren Tools für das Vorfallmanagement zählt Automatisierung, die Kommunikationsstreams wie Slack, Zoom und Confluence doc erstellt. Es werden jedoch keine untergeordneten Streams erstellt, die bei Vorfällen großer Größenordnung erforderlich sind, um Wiederherstellungsstreams zu isolieren.

Maßnahmen:
Definieren eines Playbooks und von Tools für Vorfälle großer Größenordnung und Ausführen simulierter Übungen: Wir definieren und dokumentieren die Vorfälle, die als schwerwiegend angesehen werden und dieses Maß an Reaktion erfordern. Wir beschreiben wichtige Koordinationsschritte und entwickeln Tools, die Incident Managern und anderen Geschäftsfunktionen dabei helfen, die Reaktion zu optimieren und die Wiederherstellung zu starten Incident Manager führen mit ihren Teams regelmäßig Simulationen, Schulungen und Verbesserungen der Tools und Dokumentation durch, um sie kontinuierlich zu optimieren.

Erkenntnis 4: Verbesserung unserer Kommunikationsprozesse

a) Wir haben kritische Kunden-IDs gelöscht, was sich auf die Kommunikation und Maßnahmen in Bezug auf die Betroffenen auswirkte
Das gleiche Skript, mit dem Kunden-Sites gelöscht wurden, löschte auch wichtige Kunden-IDs (z. B. Site-URL, Site-Systemadministrator-Kontakte) aus unseren Produktionsumgebungen. Infolgedessen wurden (1) Kunden daran gehindert, Tickets für technischen Support über unseren normalen Support-Kanal einzureichen; (2) es dauerte mehrere Tage, bis wir eine zuverlässige Liste der wichtigsten Kundenkontakte (wie Systemadministratoren der Site) erhielten, die von dem Ausfall betroffen waren, um proaktiv agieren zu können und (3) Support-Workflows, SLAs, Dashboards und Eskalationsprozesse funktionierten anfangs aufgrund der Einzigartigkeit des Vorfalls nicht ordnungsgemäß.

Während des Ausfalls erfolgten Kundeneskalationen auch über mehrere Kanäle (E-Mail, Telefonanrufe, CEO-Tickets, LinkedIn und andere soziale Kanäle sowie Support-Tickets). Unterschiedliche Tools und Prozesse in unseren Teams mit Kundenkontakt verlangsamten unsere Reaktion und erschwerten die ganzheitliche Verfolgung und Berichterstattung zu diesen Eskalationen.

b) Wir hatten kein Playbook für die Kommunikation bei Vorfällen, das detailliert genug war, um dieser Komplexität gerecht zu werden

Wir hatten kein Playbook für die Kommunikation von Vorfällen, in dem Prinzipien sowie Rollen und Verantwortlichkeiten beschrieben wurden, um schnell genug ein einheitliches, funktionsübergreifendes Team für die Kommunikation bei einem Vorfall zu mobilisieren. Wir haben den Vorfall nicht schnell und konsistent über mehrere Kanäle, insbesondere in sozialen Medien, bestätigt. Eine umfassendere öffentliche Kommunikation im Zusammenhang mit dem Ausfall sowie die Wiederholung der wichtigen Botschaft, dass es keinen Datenverlust gab und dies nicht das Ergebnis eines Cyberangriffs war, wäre der richtige Ansatz gewesen.

Maßnahmen:

- **Verbesserung der Sicherung wichtiger Kontakte:** Sicherung autorisierter Kontokontaktinformationen außerhalb der Produktinstanz.
- **Nachrüsten von Support-Tools:** Erstellen von Mechanismen für Kunden ohne gültige Site-URL oder Atlassian-ID, um direkten Kontakt mit unserem technischen Support-Team aufzunehmen.
- **Kundeneskalationssystem und -prozesse:** Investition in ein einheitliches, kontobasiertes Eskalationssystem und Workflows, die es ermöglichen, mehrere Arbeitsobjekte (Tickets, Aufgaben usw.) unter einem einzigen Kundenkonto-Objekt zu speichern, um eine bessere Koordination und Transparenz in allen unseren Teams mit Kundenkontakt zu erzielen.
- **Schnellere Abdeckung des Eskalationsmanagements rund um die Uhr:** Anpassung an globale Erweiterungspläne für das Eskalationsmanagement, um eine Abdeckung durch erfahrene Mitarbeiter in allen großen Regionen rund um die Uhr zu gewährleisten, gemeinsam mit Support-Rollen, die Unterstützung durch Produkt- und Vertriebsexperten und Führungskräfte bieten.
- **Aktualisieren unseres Playbooks für die Kommunikation bei Vorfällen mit neuen Erkenntnissen und regelmäßige Überprüfung:** Überprüfung des Playbooks, um intern klare Rollen und Kommunikationswege zu definieren. Nutzung des DACI-Frameworks für Vorfälle und Festlegen von Mitarbeitern, die bei Krankheit, an Feiertagen oder bei anderen unvorhersehbaren Ereignissen jederzeit einspringen können. Durchführung einer vierteljährlichen Überprüfung, um jederzeit die Einsatzbereitschaft zu überprüfen.

Maßnahmen (Fortsetzung)
Einhaltung der Vorlage für die Kommunikation bei Vorfällen: Wir gehen darauf ein, was passiert ist, wer betroffen war, wie die Zeitleiste für die Wiederherstellung aussieht, wie viel Prozent der Sites wiederhergestellt wurden, wie hoch der erwartete Datenverlust ist und wie zuversichtlich das Unternehmen bezüglich der Wiederherstellung ist. Wir stellen außerdem eine klare Anleitung für den Kontakt des Kundensupports zur Verfügung.

Abschließende Bemerkungen

Der Ausfall ist behoben und die Kunden-Sites sind vollständig wiederhergestellt, doch unsere Arbeit geht weiter. In dieser Phase implementieren wir die oben beschriebenen Änderungen, um unsere Prozesse zu verbessern, unsere Ausfallsicherheit zu erhöhen und zu verhindern, dass sich eine Situation wie diese wiederholt.

Atlassian hört nie auf, dazuzulernen, und unsere Teams haben aus dieser Erfahrung sicherlich viele schwierige Lektionen gelernt. Wir setzen diese Lektionen in die Tat um, um unser Geschäft nachhaltig zu verändern. Letztlich werden wir stärker daraus hervorgehen und Ihnen aufgrund dieser Erfahrung einen besseren Service bieten.

Wir hoffen, dass die Erkenntnisse aus diesem Vorfall anderen Teams helfen werden, die fleißig daran arbeiten, ihren Kunden zuverlässige Services zu bieten.

Abschließend möchte ich allen danken, die dies lesen und mit uns lernen, sowie denen, die Teil unserer erweiterten Atlassian Community und unseres Teams sind.

– Sri Viswanath, CTO