ATLASSIAN

Revisión posterior al incidente

Interrupción del servicio de abril de 2022

Esta traducción solo pretende servir de referencia. Si hubiera alguna ambigüedad o contradicción entre las versiones traducidas, prevalecerá la versión en inglés.

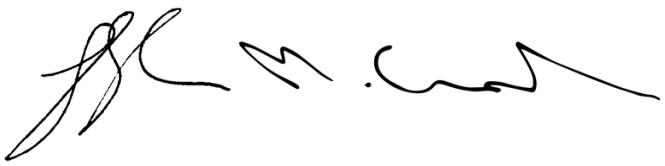
Carta de nuestros cofundadores y codirectores ejecutivos

Queremos reconocer la interrupción del servicio que afectó a los clientes a principios de este mes. Somos conscientes de que nuestros productos son fundamentales para tu negocio y no nos tomamos esta responsabilidad a la ligera. Es responsabilidad nuestra. Fin de la historia. Para los clientes afectados, estamos trabajando para recuperar vuestra confianza.

> En Atlassian, uno de nuestros valores fundamentales es «Empresa abierta, sin tonterías». En parte, ponemos en práctica este valor hablando abiertamente sobre los incidentes y usándolos como oportunidades para aprender. Esta revisión posterior al incidente está dirigida a nuestros clientes, a nuestra comunidad de Atlassian y a la comunidad técnica en general. En Atlassian estamos orgullosos de nuestro proceso de gestión de incidentes, que pone de relieve que una política corporativa en la que no se buscan culpables y centrarnos en identificar formas de mejorar nuestros sistemas y procesos técnicos es fundamental para ofrecer servicios fiables y de gran escala. Si bien hacemos todo lo posible para evitar cualquier tipo de incidente, también aceptamos la idea de que los incidentes son una forma eficaz de mejorar.

Ten por seguro que la plataforma en la nube de Atlassian nos permite satisfacer las diversas necesidades de nuestros más de 200 000 clientes en la nube de todos los tamaños y sectores. Antes de este incidente, nuestra nube ofrecía sistemáticamente acuerdos de nivel de servicio (SLA) con un 99,9 % de tiempo de actividad y con un tiempo de actividad superior. Hemos realizado inversiones a largo plazo en nuestra plataforma y en varias funciones de plataforma centralizada, con una infraestructura escalable y una cadencia constante de mejoras de seguridad.

A nuestros clientes y socios, os damos las gracias por vuestra confianza y colaboración continuas. Esperamos que los detalles y las acciones que se describen en este documento demuestren que Atlassian seguirá ofreciendo una plataforma en la nube de primera calidad y una potente cartera de productos para satisfacer las necesidades de cada equipo.



-Scott y Mike

Resumen ejecutivo

El martes 5 de abril de 2022, a partir de las 7:38 UTC, 775 clientes de Atlassian perdieron el acceso a sus productos de Atlassian. La interrupción duró hasta 14 días para un subconjunto de estos clientes; el servicio del primer grupo de clientes se restauró el 8 de abril y todos los sitios de los clientes se restauraron el 18 de abril.

Esto no fue el resultado de un ciberataque y no hubo ningún acceso no autorizado a los datos de los clientes. Atlassian cuenta con un programa integral de gestión de datos con SLA publicados y un historial de superación de dichos SLA.

Aunque fue un incidente grave, ningún cliente perdió más de cinco minutos de datos. Además, más del 99,6 % de nuestros clientes y usuarios continuaron utilizando nuestros productos en la nube sin ninguna interrupción durante las actividades de restauración.

> A lo largo de este documento, nos referimos a los clientes cuyos sitios se eliminaron a causa de este incidente como clientes «afectados». En esta revisión posterior al incidente (PIR), se proporcionan los detalles exactos del incidente, se exponen los pasos que tomamos para la recuperación y se describe cómo evitaremos que situaciones como esta ocurran en el futuro. Proporcionamos un resumen general del incidente en esta sección, con más detalles en el resto del documento.

¿Qué sucedió?

En 2021, completamos la adquisición e integración de una aplicación de Atlassian independiente para Jira Service Management y Jira Software llamada «Insight – Asset Management». En ese momento, la funcionalidad de esta aplicación independiente era nativa en Jira Service Management y ya no estaba disponible para Jira Software. Por este motivo, necesitábamos eliminar la antigua aplicación independiente en los sitios de clientes que la tenían instalada. Nuestros equipos de ingeniería utilizaron un script y un proceso existentes para eliminar instancias de esta aplicación independiente, pero hubo dos problemas:

- **Brecha de comunicación.** En primer lugar, hubo una brecha de comunicación entre el equipo que solicitó la eliminación y el equipo que la ejecutó. En lugar de

proporcionar los ID de la *aplicación* que se había marcado para su eliminación, el equipo proporcionó los ID de *todo el sitio en la nube* en el que se iban a eliminar las aplicaciones.

- **Advertencias del sistema insuficientes.** En segundo lugar, la API utilizada para realizar la eliminación aceptaba identificadores de sitio y de aplicación y supuso que la entrada era correcta; esto significaba que, si se pasaba el ID de un sitio, se eliminaba un sitio; si se pasaba el ID de una aplicación, se eliminaba una aplicación. No hubo ninguna señal de advertencia para confirmar el tipo de eliminación (sitio o aplicación) que se estaba solicitando.

El script que se ejecutó siguió nuestro proceso estándar de revisión por pares, que se centró en qué punto final se llamaba y cómo. No se realizaron comprobaciones adicionales de los ID de sitios en la nube proporcionados para validar si hacían referencia a la Insight App o a todo el sitio, y el problema fue que el script contenía el ID de todo el sitio de un cliente. El resultado fue la eliminación inmediata de 883 sitios (que representan a 775 clientes) entre las 07:38 UTC y las 08:01 UTC del martes 5 de abril de 2022. *Consulta «Qué pasó»*

¿Cómo respondimos?

Una vez que se confirmó el incidente el 5 de abril a las 08:17 UTC, iniciamos nuestro proceso de gestión de incidentes graves y formamos un equipo de gestión de incidentes multifuncional. El equipo global de respuesta ante incidentes trabajó ininterrumpidamente durante el incidente hasta que todos los sitios fueron restaurados, validados y devueltos a los clientes. Además, los líderes de gestión de incidentes se reunieron cada tres horas para coordinar los flujos de trabajo.

Al principio, nos dimos cuenta de que restaurar a cientos de clientes con varios productos simultáneamente comportaba una serie de desafíos.

Cuando empezó el incidente, sabíamos exactamente qué sitios estaban afectados y nuestra prioridad era establecer comunicación con el propietario aprobado de cada sitio afectado para informarle de la interrupción.

Sin embargo, se eliminó parte de la información de contacto de los clientes. Esto significaba que los clientes no podían presentar tickets de asistencia como lo harían normalmente. También significaba que no teníamos acceso inmediato a los contactos clave de los clientes. *Para obtener más información, consulta «Resumen general de los flujos de trabajo de recuperación»*

¿Qué estamos haciendo para evitar situaciones como esta en el futuro?

Hemos tomado una serie de medidas inmediatas y nos comprometemos a realizar cambios para evitar esta situación en el futuro. Estas son cuatro áreas específicas en las que hemos realizado o realizaremos cambios significativos:

1. **Establecer «borrados suaves» universales en todos los sistemas**. En general, una eliminación de este tipo debe estar prohibida o tener múltiples capas de protección para evitar errores, incluyendo un plan de implementación escalonada y reversión probada de «borrados suaves». Evitaremos globalmente la eliminación de los datos y metadatos de clientes que no hayan pasado por un proceso de borrado suave.

2. **Invertir en nuestro programa de recuperación ante desastres (DR) para automatizar la restauración en caso de eliminación de múltiples sitios y productos para un conjunto más grande de clientes.** Aprovecharemos la automatización y los aprendizajes de este incidente para acelerar el programa de DR y cumplir con el objetivo de tiempo de recuperación (RTO), tal como se define en nuestra política para esta magnitud de incidente. Realizaremos regularmente ejercicios de DR que implican la restauración de todos los productos para un gran conjunto de sitios.

3. **Mejorar el proceso de gestión de incidentes para incidentes a gran escala**. Mejoraremos nuestro procedimiento operativo estándar para incidentes a gran escala y practicaremos con simulaciones de esta magnitud de incidente. Actualizaremos nuestra formación y herramientas para gestionar el gran número de equipos que trabajan en paralelo.

4. **Crear un manual de estrategias para incidentes a gran escala**. Admitiremos los incidentes de manera temprana a través de múltiples canales. Difundiremos comunicaciones públicas sobre los incidentes en cuestión de horas. Para llegar mejor a los clientes afectados, mejoraremos la copia de seguridad de los contactos clave y actualizaremos las herramientas de soporte para que los clientes que no tengan una URL válida o un ID de Atlassian puedan ponerse en contacto directamente con nuestro equipo de asistencia técnica.

Nuestra lista completa de elementos de acción se detalla en la revisión completa posincidente que se incluye a continuación. *Consulta «Cómo mejoraremos»*

Índice

Descripción general de la arquitectura en la nube de Atlassian Página 7
- Arquitectura de alojamiento en la nube de Atlassian
- Arquitectura de los servicios distribuidos
- Arquitectura de varios inquilinos
- Aprovisionamiento y ciclo de vida de inquilinos
- Programa de recuperación ante desastres
    - Resistencia
    - Capacidad de restauración del almacenamiento de servicios
    - Capacidad de restauración automatizada de múltiples sitios y productos

Qué sucedió, cronograma y recuperación Página 13
- Qué pasó
- Cómo nos coordinamos
- Cronograma del incidente
- Resumen general de los flujos de trabajo de la recuperación
    - Flujo de trabajo 1: detección, inicio de la recuperación e identificación de nuestro enfoque
    - Flujo de trabajo 2: recuperación temprana y el enfoque «Restauración 1»
    - Flujo de trabajo 3: recuperación acelerada y el enfoque «Restauración 2»
    - Pérdida mínima de datos tras la restauración de sitios eliminados

Comunicación de incidentes Página 21
- Qué pasó

Experiencia de asistencia y contacto con el cliente Página 23
- ¿Cómo se vio afectada la asistencia para nuestros clientes?
- ¿Cómo respondimos?

¿Cómo mejoraremos? Página 25
- Aprendizaje 1: los «borrados suaves» deben ser universales en todos los sistemas
- Aprendizaje 2: como parte del programa de DR, automatizar la restauración de eventos de eliminación de múltiples sitios y productos para un conjunto más grande de clientes
- Aprendizaje 3: mejorar el proceso de gestión de incidentes para eventos a gran escala
- Aprendizaje 4: mejorar nuestros procesos de comunicación

Observaciones finales Página 31

Descripción general de la arquitectura en la nube de Atlassian

Para comprender los factores que contribuyeron al incidente, tal como se explica en este documento, es útil entender primero la arquitectura de implementación de los productos, servicios e infraestructura de Atlassian.

Arquitectura de alojamiento en la nube de Atlassian

Atlassian utiliza Amazon Web Services (AWS) como proveedor de servicios en la nube y sus instalaciones de centros de datos de gran disponibilidad en varias regiones del mundo. Cada región de AWS es una ubicación geográfica independiente con diversos grupos de centros de datos aislados y físicamente separados, conocidos como zonas de disponibilidad (AZ).

Aprovechamos los servicios informáticos, de almacenamiento, de red y de datos de AWS para generar nuestros productos y componentes de plataforma, lo que nos permite utilizar las capacidades de redundancia que ofrece AWS, como las zonas y regiones de disponibilidad.

Arquitectura de los servicios distribuidos

Con esta arquitectura de AWS, alojamos una serie de servicios de plataforma y de producto que se utilizan en nuestras soluciones. Esto incluye funciones de plataforma que se comparten y consumen en varios productos de Atlassian, como Media, Identity, Commerce, experiencias como nuestro Editor, así como funciones específicas de productos, como el servicio Jira Issue y el Análisis de Confluence.

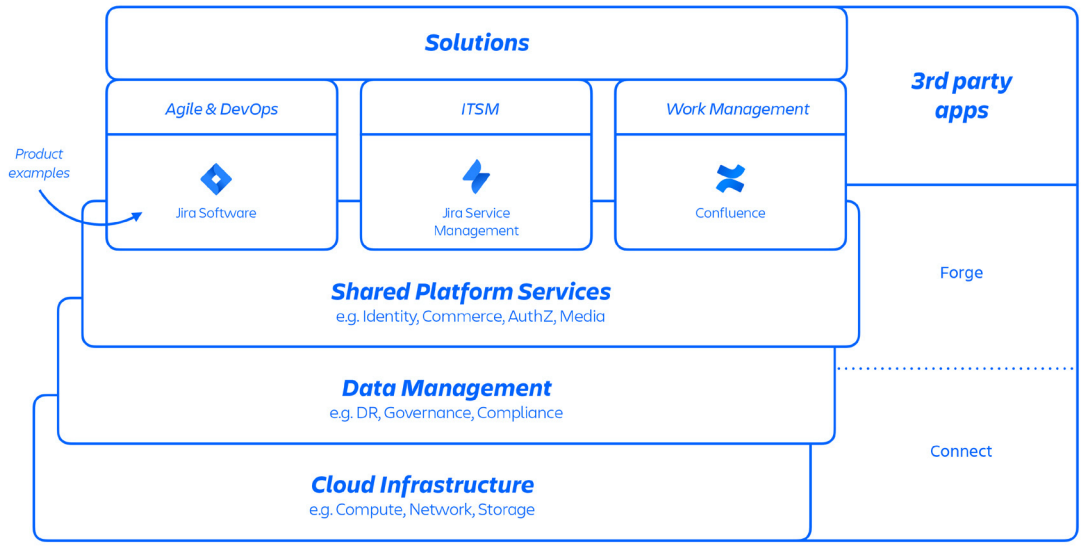


Figura 1: arquitectura de la plataforma de Atlassian.

Los desarrolladores de Atlassian proporcionan estos servicios a través de una plataforma como servicio (PaaS) desarrollada internamente, llamada Micros, que orquesta automáticamente la implementación de servicios compartidos, infraestructura, almacenes de datos y sus capacidades de gestión, incluidos los requisitos de control de seguridad y cumplimiento (consulta la *figura 1* arriba). Por lo general, un producto de Atlassian consiste en varios servicios «en contenedores» que se implementan en AWS mediante Micros. Los productos de Atlassian utilizan funciones principales de la plataforma (consulta la *figura 2* a continuación) que van desde el enrutamiento de solicitudes a los almacenes de objetos binarios, la autenticación o autorización, el contenido generado por los usuarios (UGC) transaccional y los almacenes de relaciones de entidades, los lagos de datos, el registro común, el seguimiento de solicitudes, la observabilidad y los servicios de análisis. Estos microservicios se generan utilizando recursos técnicos aprobados y estandarizados a nivel de plataforma:

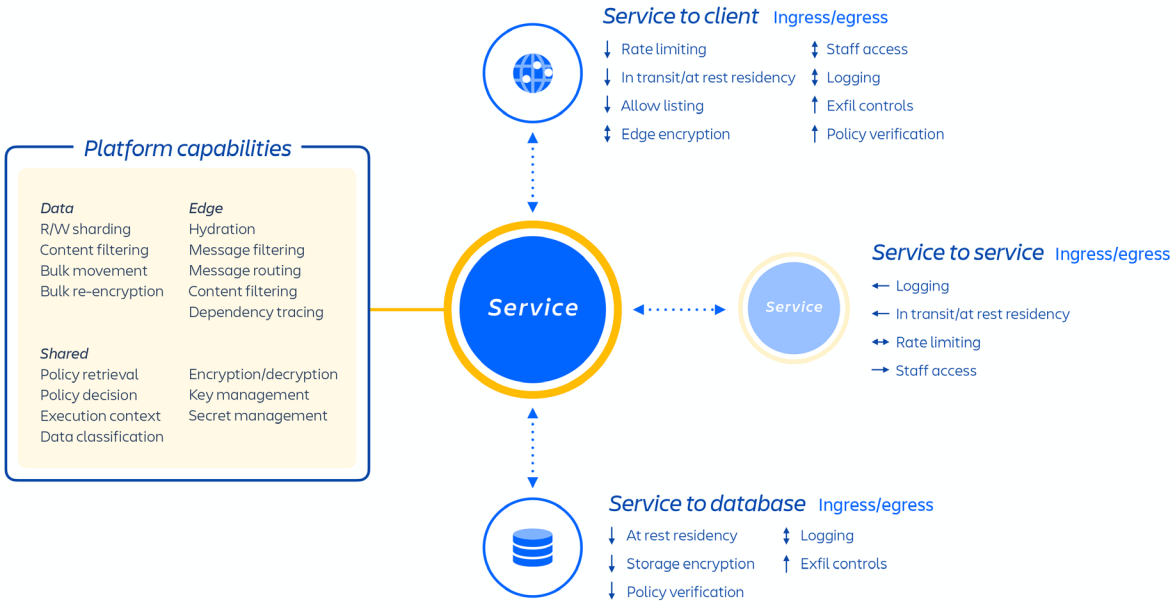


Figura 2: descripción general de los microservicios de Atlassian.

Arquitectura de varios inquilinos

Además de nuestra infraestructura en la nube, hemos desarrollado y utilizamos una arquitectura de microservicios de varios inquilinos junto con una plataforma compartida que respalda nuestros productos. En la arquitectura de varios inquilinos, un solo servicio sirve a varios clientes, incluyendo las bases de datos y las instancias de cómputo necesarias para ejecutar nuestros productos en la nube. Cada partición (esencialmente un contenedor; consulta la *figura 3* a continuación) contiene los datos de varios inquilinos, pero los datos de cada inquilino están aislados y son inaccesibles para otros inquilinos.

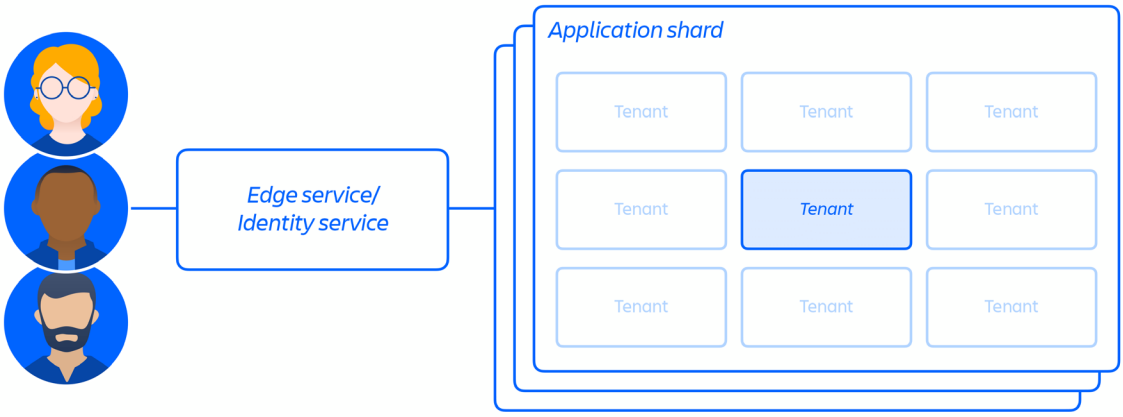


Figura 3: cómo almacenamos datos en una arquitectura de varios inquilinos.

Aprovisionamiento y ciclo de vida de inquilinos

Cuando se aprovisiona a un cliente nuevo, una serie de eventos desencadenan la orquestación de servicios distribuidos y el aprovisionamiento de almacenes de datos. Por lo general, estos eventos se pueden asignar a uno de los siete pasos del ciclo de vida:

1. Los sistemas de comercio se actualizan inmediatamente con los metadatos y la información de control de acceso más recientes para ese cliente y, a continuación, un sistema de orquestación de aprovisionamiento alinea el «estado de los recursos aprovisionados» con el estado de la licencia a través de una serie de eventos de inquilinos y productos.

Eventos para inquilinos
Estos eventos afectan al inquilino en su conjunto y pueden ser:

- Creación: se crea un inquilino y se utiliza para sitios nuevos
- Destrucción: se elimina todo un inquilino

Eventos de productos

- Activación: tras la activación de productos con licencia o aplicaciones de terceros
- Desactivación: tras la desactivación de determinados productos o aplicaciones
- Suspensión: tras la suspensión de un producto existente determinado, deshabilitando así el acceso de los clientes a un sitio del que son propietarios
- Anulación de la suspensión: tras la anulación de la suspensión de un producto existente determinado, habilitando así el acceso de los clientes a un sitio del que son propietarios

Actualización de licencia: contiene información sobre el número de licencias de un producto determinado, así como sobre su estado (activo/inactivo)

2. Creación del sitio del cliente y activación del conjunto correcto de productos para el cliente. El concepto de un sitio es el contenedor de varios productos con licencia para un cliente en particular (por ejemplo, Confluence y Jira Software para <site-name>.atlassian.net). Esto (consulta la *figura 4* a continuación) es un punto importante que entender en el contexto de este informe, ya que el contenedor del sitio es lo que se eliminó en este incidente, y el concepto del sitio se aborda a lo largo del documento.

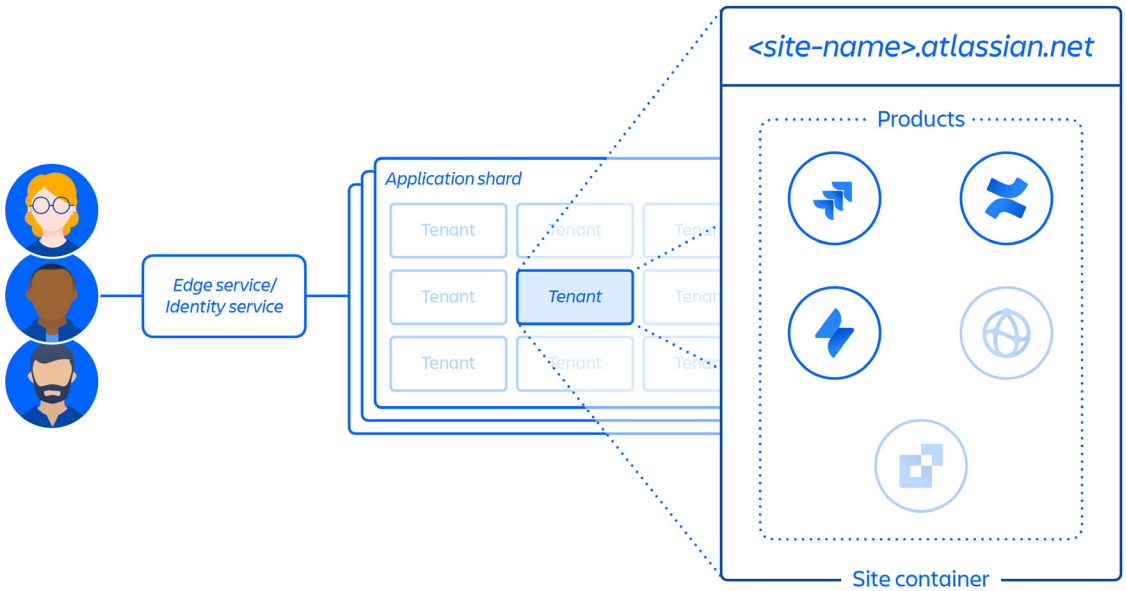


Figura 4: descripción general del contenedor del sitio.

3. Aprovisionamiento de productos dentro del sitio del cliente en la región designada.

   Cuando se aprovisiona un producto, la mayoría de su contenido estará alojado cerca de donde los usuarios acceden a él. Para optimizar el rendimiento del producto, no limitamos el movimiento de datos cuando están alojados en todo el mundo y podemos mover datos entre regiones según sea necesario.

   Para algunos de nuestros productos, también ofrecemos un servicio de residencia de datos. La residencia de datos permite a los clientes elegir si los datos de los productos se distribuyen globalmente o se mantienen en una de nuestras ubicaciones geográficas definidas.

4. Creación y almacenamiento de los metadatos principales y de la configuración de los productos y del sitio del cliente.

5. Creación y almacenamiento de los datos de identidad del sitio y de los productos, como usuarios, grupos, permisos, etc.

6. Aprovisionamiento de bases de datos de productos en un sitio, p. ej. la familia de productos Jira, Confluence, Compass y Atlas.

7. Aprovisionamiento de las aplicaciones con licencia de los productos.

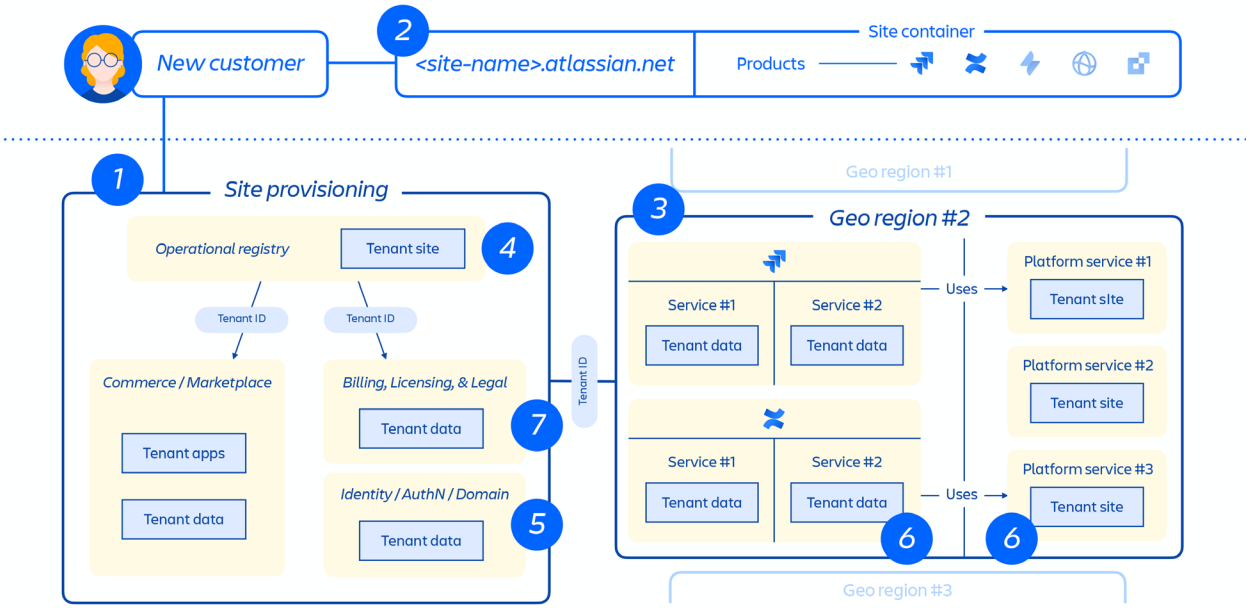


Figura 5: descripción general de cómo se aprovisiona el sitio del cliente en toda nuestra arquitectura distribuida.

La *figura 5* que hay arriba demuestra cómo se implementa el sitio de un cliente en nuestra arquitectura distribuida, no solo en una única base de datos o almacén. Esto incluye varias ubicaciones físicas y lógicas que almacenan metadatos, datos de configuración, datos de productos, datos de plataformas y otra información del sitio relacionada.

Programa de recuperación ante desastres

Nuestro programa de recuperación ante desastres (DR) engloba todos nuestros esfuerzos para brindar resiliencia contra fallos de infraestructura y capacidad de restauración del almacenamiento de servicios a partir de copias de seguridad. Dos conceptos importantes para entender los programas de recuperación ante desastres son los siguientes:

- **Objetivo de tiempo de recuperación (RTO):** ¿con qué rapidez se pueden recuperar los datos y devolverlos a un cliente durante un desastre?
- **Objetivo de punto de recuperación (RPO):** ¿hasta qué punto están actualizados los datos recuperados después de recuperarlos a partir de una copia de seguridad? ¿Cuántos datos se perderán desde la última copia de seguridad?

Durante este incidente, no alcanzamos nuestro RTO, pero cumplimos con nuestro RPO.

Resistencia

Nos preparamos para los fallos a nivel de infraestructura; por ejemplo, la pérdida de toda una base de datos, de un servicio o de zonas de disponibilidad de AWS. Esta preparación incluye la réplica de datos y servicios en varias zonas de disponibilidad y pruebas de conmutación por error periódicas.

Capacidad de restauración del almacenamiento de servicios

También nos preparamos para recuperarnos de la corrupción de datos del almacenamiento de servicios debido a riesgos como ransomware, actores maliciosos, defectos de software y errores operativos. Esta preparación incluye copias de seguridad inmutables y pruebas de restauración de copias de seguridad de almacenamiento de servicio. Podemos tomar cualquier almacén de datos individual y restaurarlo a un punto anterior en el tiempo.

Capacidad de restauración automatizada de múltiples sitios y productos

En el momento del incidente, no teníamos la capacidad de seleccionar un conjunto grande de sitios de clientes y restaurar todos sus productos interconectados desde copias de seguridad hasta un punto anterior en el tiempo.

Nuestras capacidades se han centrado en la infraestructura, la corrupción de datos, los eventos de servicio único o las eliminaciones de un solo sitio. En el pasado, hemos tenido que lidiar con este tipo de fallos y probarlos. La eliminación a nivel de sitio no tenía runbooks que pudieran automatizarse rápidamente para la escala de este evento, que requería que las herramientas y la automatización en todos los productos y servicios se llevaran a cabo de manera coordinada.

En las siguientes secciones se profundiza en esta complejidad y en lo que estamos haciendo en Atlassian para evolucionar y optimizar nuestras capacidades a fin de mantener esta arquitectura a escala.

Qué sucedió, cronograma y recuperación

Qué pasó

En 2021, completamos la integración de una aplicación de Atlassian independiente para Jira Service Management y Jira Software, llamada «Insight – Asset Management». En ese momento, la funcionalidad de esta aplicación independiente era nativa en Jira Service Management y ya no estaba disponible para Jira Software. Por este motivo, necesitábamos eliminar la antigua aplicación independiente en los sitios de clientes que la tenían instalada. Nuestros equipos de ingeniería utilizaron un script y un proceso existentes para eliminar instancias de esta aplicación independiente.

Sin embargo, se produjeron dos problemas críticos:

- **Brecha de comunicación.** En primer lugar, hubo una brecha de comunicación entre el equipo que solicitó la eliminación y el equipo que la ejecutó. En lugar de proporcionar los ID de la aplicación que se había marcado para su eliminación, el equipo proporcionó los ID de todo el sitio en la nube en el que se iban a desactivar las aplicaciones.
- **Advertencias del sistema insuficientes.** En segundo lugar, la API utilizada para realizar la eliminación aceptaba identificadores de sitio y de aplicación y supuso que la entrada era correcta; esto significaba que, si se pasaba el ID de un sitio, se eliminaba un sitio; si se pasaba el ID de una aplicación, se eliminaba una aplicación. No había ninguna señal de advertencia para confirmar el tipo de eliminación (sitio o aplicación) que se estaba solicitando.

El script que se ejecutó siguió nuestro proceso estándar de revisión por pares, que se centró en qué punto final se llamaba y cómo. No se realizaron comprobaciones adicionales de los ID de sitios en la nube proporcionados para validar si hacían referencia a la aplicación o a todo el sitio. El script se probó en un entorno de ensayo según nuestros procesos de gestión de cambios estándar; sin embargo, no habría detectado que la entrada de los ID fuera incorrecta, ya que los ID no existían en el entorno de ensayo.

Cuando se ejecutó en producción, el script se ejecutó inicialmente en 30 sitios. La primera ejecución en producción fue un éxito y eliminó la aplicación Insight de esos 30 sitios sin otros efectos secundarios. Sin embargo, los ID de esos 30 sitios se obtuvieron antes del evento de falta de comunicación e incluían los ID correctos de la aplicación Insight.

Sin embargo, el script para la posterior ejecución en producción incluía ID de sitios en lugar de ID de la aplicación Insight y se ejecutó en un conjunto de 883 sitios. El script

comenzó a ejecutarse el 5 de abril a las 07:38 UTC y se completó a las 08:01 UTC. El script eliminó sitios secuencialmente en función de la lista de entrada, por lo que el sitio del primer cliente se eliminó poco después de que el script comenzara a ejecutarse a las 07:38 UTC. El resultado fue la eliminación inmediata de los 883 sitios, sin señal de advertencia para nuestros equipos de ingeniería.

Los siguientes productos de Atlassian no estaban disponibles para los clientes afectados: la familia de productos Jira, Confluence, Atlassian Access, Opsgenie y Statuspage.

En cuanto nos enteramos del incidente, nuestros equipos se centraron en la restauración del servicio para todos los clientes afectados. En ese momento, calculamos que el número de sitios afectados era de unos 700 (se vieron afectados un total de 883 sitios, pero restamos los sitios que pertenecen a Atlassian). De los 700 casos, una parte significativa eran cuentas inactivas, gratuitas o pequeñas con un número bajo de usuarios activos. En base a esto, inicialmente calculamos que el número aproximado de clientes afectados era de unos 400.

Ahora tenemos una visión mucho más precisa y, para ofrecer una transparencia total en base a los registros oficiales para clientes de Atlassian, 775 clientes se vieron afectados por la interrupción. Sin embargo, la mayoría de los usuarios estaban representados dentro de la estimación original de 400 clientes. La interrupción duró hasta 14 días para un subconjunto de estos clientes; el servicio para el primer grupo de clientes se restauró el 8 de abril, y el 18 de abril ya se había restaurado el de todos los clientes.

Cómo nos coordinamos

El primer ticket de asistencia lo creó un cliente afectado a las 07:46 UTC del 5 de abril. Nuestra supervisión interna no detectó ninguna incidencia porque los sitios se eliminaron mediante un flujo de trabajo estándar. A las 08:17 UTC, iniciamos nuestro proceso de gestión de incidentes graves, formamos un equipo de gestión de incidentes multifuncional y, en siete minutos, a las 08:24 UTC, se escaló a Crítico. A las 08:53 UTC, nuestro equipo confirmó que el ticket de asistencia del cliente y la ejecución del script estaban relacionados. Una vez que nos dimos cuenta de la complejidad de la restauración, asignamos nuestro nivel más alto de gravedad al incidente a las 12:38 UTC.

El equipo de gestión de incidentes estaba compuesto por personas de varios equipos de Atlassian, incluidos los de ingeniería, atención al cliente, gestión de programas, comunicaciones y muchos más. El equipo principal se reunió cada tres horas durante el incidente hasta que todos los sitios se restauraron, se validaron y se devolvieron a los clientes.

Para gestionar el progreso de restauración, creamos un nuevo proyecto de Jira llamado SITE y un flujo de trabajo para hacer un seguimiento de las restauraciones sitio por sitio en varios equipos (ingeniería, gestión de programas, soporte, etc.). Este enfoque permitió que todos los equipos pudieran identificar fácilmente los problemas relacionados con la restauración de cualquier sitio individual y hacer un seguimiento.

También implementamos una congelación del código en toda la ingeniería durante el incidente del 8 de abril a las 03:30 UTC. Esto nos permitió centrarnos en la restauración del servicio para los clientes, eliminar el riesgo de que el cambio provocara inconsistencias en los datos de los clientes, minimizar el riesgo de otras interrupciones y reducir la probabilidad de que los cambios no relacionados distrajeran al equipo de la recuperación.

Cronograma del incidente

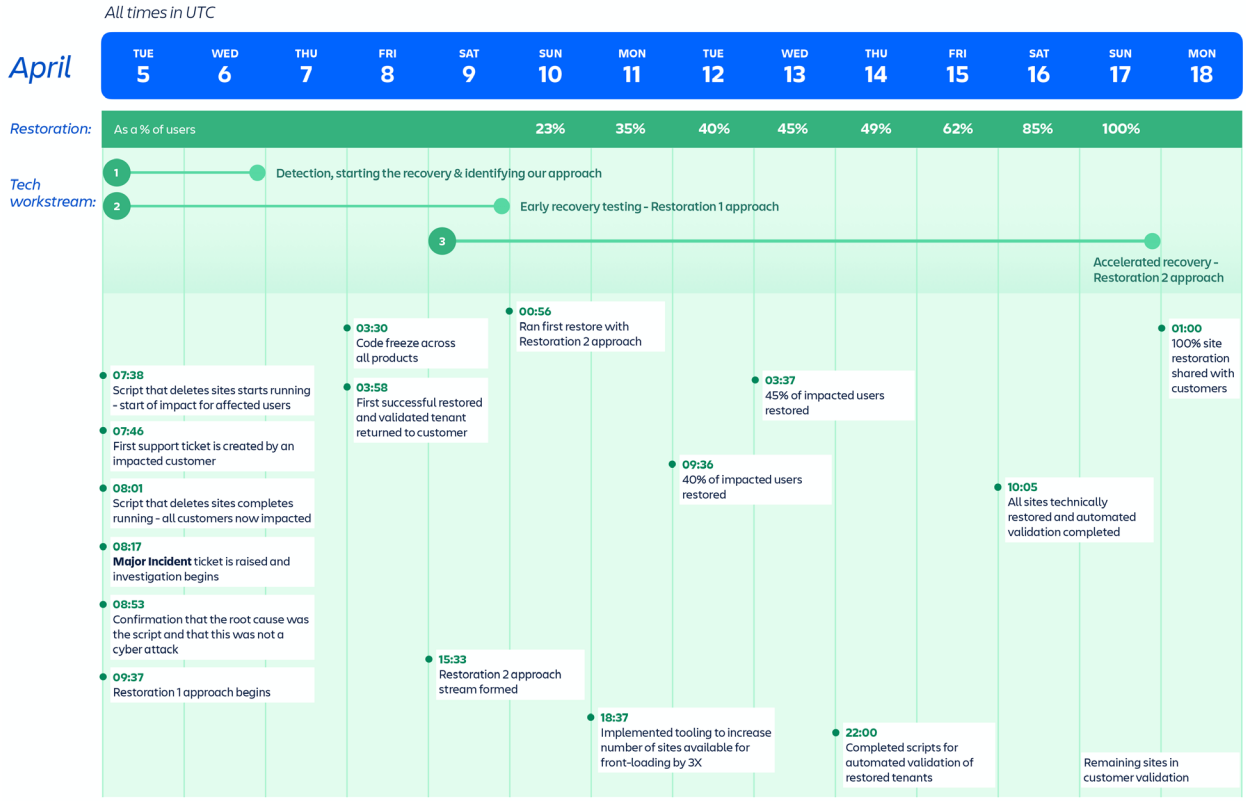


Figura 6: cronograma del incidente e hitos clave de la restauración.

Resumen general de los flujos de trabajo de la recuperación

La recuperación se ejecutó mediante tres flujos de trabajo principales: detección, recuperación temprana y aceleración. Aunque a continuación describimos cada flujo de trabajo por separado, durante la recuperación se trabajó en paralelo en todos los flujos de trabajo.

Flujo de trabajo 1: detección, inicio de la recuperación e identificación de nuestro enfoque

Marca de tiempo: días 1 y 2 (5 y 6 de abril)
A las 08:53 UTC del 5 de abril, identificamos que el script de la aplicación Insight había causado la eliminación de sitios. Confirmamos que esto no fue el resultado de un acto malicioso interno o de un ciberataque. Se llamó a los equipos de infraestructura de productos y plataformas relevantes y se les incorporó al incidente.

Al principio del incidente, reconocimos lo siguiente:

- La restauración de cientos de sitios eliminados es un proceso complejo de varios pasos (detallados en la sección de arquitectura anterior), que requiere muchos equipos y varios días para completarse correctamente.
- Podíamos recuperar un solo sitio, pero no habíamos generado capacidades ni procesos para recuperar un lote grande de sitios.

Como resultado, necesitábamos paralelizar y automatizar sustancialmente el proceso de restauración para ayudar a los clientes afectados a recuperar el acceso a sus productos de Atlassian lo antes posible.

El flujo de trabajo 1 involucró a un gran número de equipos de desarrollo que participaron en las siguientes actividades:

- Identificar y ejecutar pasos de restauración para lotes de sitios en la canalización.
- Redactar y mejorar la automatización para que los equipos pudieran ejecutar pasos de restauración para un mayor número de sitios en un lote.

Flujo de trabajo 2: recuperación temprana y el enfoque «Restauración 1»

Marca de tiempo: días 1-4 (del 5 al 9 de abril)
Comprendimos qué causó la eliminación de sitios el 5 de abril a las 08:53 UTC, una hora después de que el script finalizara su ejecución. También identificamos el proceso de restauración que se había utilizado anteriormente para recuperar un pequeño número de

sitios y ponerlos en producción. Sin embargo, el proceso de recuperación para restaurar sitios eliminados a tal escala no estaba bien definido.

Para poner la recuperación en marcha rápidamente, las primeras etapas del incidente se dividieron en dos grupos de trabajo:

- El grupo de trabajo manual validó los pasos necesarios y ejecutó manualmente el proceso de restauración para un pequeño número de sitios.
- El grupo de trabajo de automatización tomó el proceso de restauración existente y generó la automatización para ejecutar los pasos de manera segura en lotes más grandes de sitios.

Descripción general del enfoque «Restauración 1» (consulta *la figura 7* a continuación):

- Este enfoque requería la creación de un sitio nuevo para cada sitio eliminado, seguido de cada producto, servicio y almacén de datos derivado que necesitara restaurar sus datos.
- El nuevo sitio contaría con nuevos identificadores, como CloudID. Todos estos identificadores se consideran inmutables, lo que significa que muchos sistemas incorporan estos identificadores en los registros de datos. Como resultado, necesitábamos actualizar grandes cantidades de datos si estos identificadores cambiaban, lo que es particularmente problemático para las aplicaciones de ecosistemas de terceros.
- La modificación de un sitio nuevo para duplicar el estado del sitio eliminado tenía dependencias complejas y a menudo imprevistas entre los pasos.

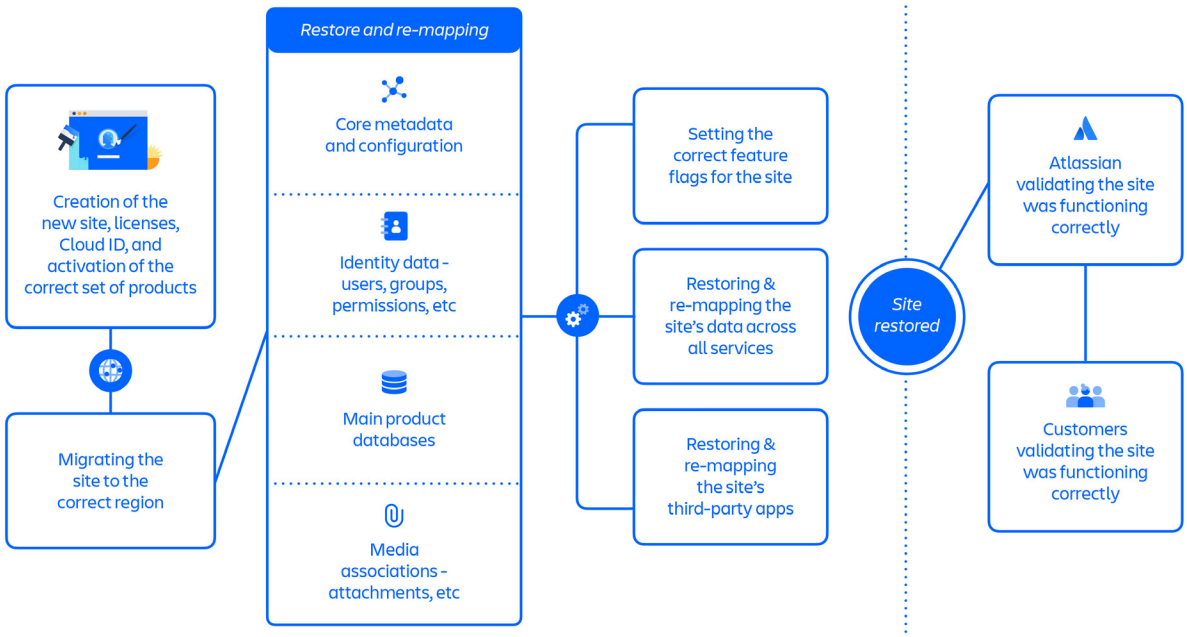


Figura 7: pasos clave del enfoque «Restauración 1».

El enfoque «Restauración 1» incluía aproximadamente 70 pasos individuales que, cuando se agregaron a un nivel alto, siguieron un flujo en gran medida secuencial de:

- Creación del nuevo sitio, licencias, Cloud ID y activación del conjunto correcto de productos
- Migración del sitio a la región correcta
- Restauración y reasignación de los metadatos principales y la configuración del sitio
- Restauración y reasignación de los datos de identidad del sitio: usuarios, grupos, permisos, etc.
- Restauración de las principales bases de datos de productos del sitio
- Restauración y reasignación de las asociaciones de medios del sitio: archivos adjuntos, etc.
- Configuración de las marcas de función correctas para el sitio
- Restauración y reasignación de los datos del sitio en todos los servicios
- Restauración y reasignación de las aplicaciones de terceros del sitio
- Validación por parte de Atlassian de que el sitio funcionaba correctamente
- Validación por parte de los clientes de que el sitio funcionaba correctamente

Una vez optimizado, el enfoque «Restauración 1» tardó aproximadamente 48 horas en restaurar un lote de sitios y se utilizó para la recuperación del 53 % de los usuarios afectados en 112 sitios entre el 5 y el 14 de abril.

Flujo de trabajo 3: recuperación acelerada y el enfoque «Restauración 2»

Marca de tiempo: días 4-13 (del 9 al 17 de abril)
Con el enfoque «Restauración 1», nos habría llevado tres semanas restaurar el servicio para todos los clientes. Por lo tanto, el 9 de abril, propusimos un nuevo enfoque para acelerar la restauración de todos los sitios, el enfoque «Restauración 2» (consulta la *figura 8* más abajo).

El enfoque «Restauración 2» ofreció un paralelismo mejorado entre los pasos de restauración al reducir la complejidad y el número de dependencias que estaban presentes con el enfoque «Restauración 1».

El proceso «Restauración 2» implicó la recreación (o la no eliminación) de los registros asociados con el sitio en todos los respectivos sistemas, empezando por el registro del Servicio de catálogo. Un elemento clave de este nuevo enfoque fue *reutilizar todos los identificadores de sitio antiguos*. Esto eliminó más de la mitad de los pasos del proceso anterior que se utilizaban para asignar los identificadores antiguos a los nuevos, incluida la necesidad de coordinarse con cada proveedor de aplicaciones de terceros para cada sitio. Sin embargo, el cambio del enfoque «Restauración 1» al enfoque «Restauración 2» agregó una sobrecarga sustancial en la respuesta ante incidentes:

- Muchos de los scripts y procesos de automatización establecidos en el enfoque «Restauración 1» tuvieron que modificarse para la «Restauración 2».
- Los equipos que realizaban restauraciones (incluidos los coordinadores de incidentes) tenían que gestionar lotes paralelos de restauraciones en ambos enfoques, mientras nosotros probábamos y validábamos el proceso «Restauración 2».
- El uso de un enfoque nuevo significaba que teníamos que probar y validar el proceso «Restauración 2» antes de ampliarlo, lo que requería duplicar el trabajo de validación que se había completado previamente para la «Restauración 1».

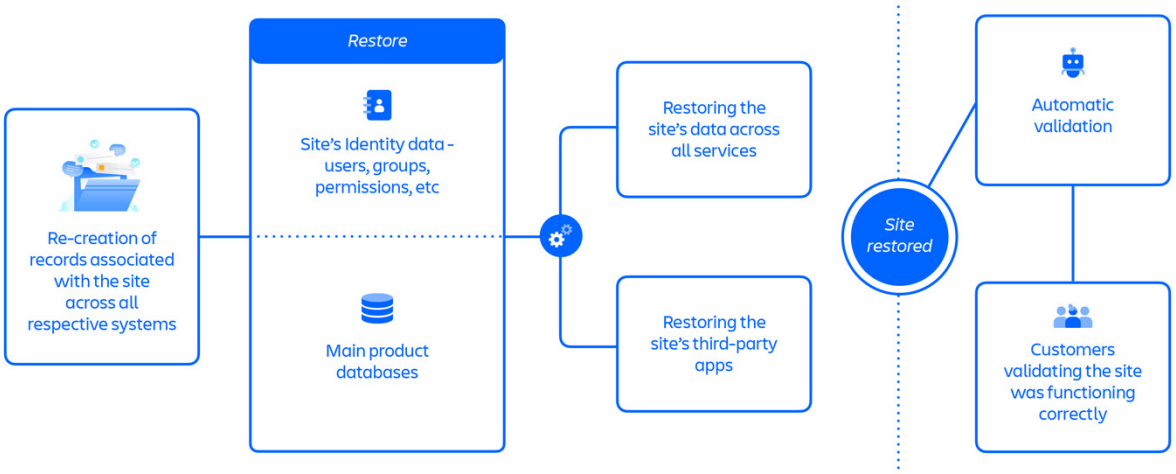


Figura 8: pasos clave del enfoque «Restauración 2».

El gráfico anterior representa el enfoque «Restauración 2», que incluía más de 30 pasos que siguieron un flujo en gran medida paralelo de:

- Recreación de registros asociados con el sitio en todos los respectivos sistemas
- Restauración de los datos de identidad del sitio: usuarios, grupos, permisos, etc.
- Restauración de las principales bases de datos de productos del sitio
- Restauración de los datos del sitio en todos los servicios
- Restauración de las aplicaciones de terceros del sitio
- Validación automática
- Validación por parte de los clientes de que el sitio funcionaba correctamente

Como parte de la recuperación acelerada, también tomamos medidas para adelantar y automatizar la restauración de sitios, ya que la restauración manual no habría escalado bien para lotes grandes. La naturaleza secuencial del proceso de recuperación significaba que la restauración del sitio podía ser más lenta para restauraciones de bases de datos grandes y restauraciones de permisos y bases de usuarios. Las optimizaciones que implementamos incluían lo siguiente:

- Desarrollamos las herramientas y los márgenes de seguridad necesarios para llevar a cabo pasos de *adelanto* y de larga duración, como restauraciones de bases de datos y sincronizaciones de identidad, para que se completaran antes que otros pasos de restauración.
- Los equipos de ingeniería generaron una automatización para los pasos individuales, lo que permitió que grandes lotes de restauraciones se ejecutaran de forma segura.
- También se generó una automatización para validar que los sitios funcionaban correctamente después de completar todos los pasos de restauración.

El enfoque de restauración acelerada «Restauración 2» tardó aproximadamente 12 horas en restaurar un sitio y se utilizó para la recuperación de aproximadamente el 47 % de los usuarios afectados en 771 sitios entre el 14 y el 17 de abril.

Pérdida mínima de datos tras la restauración de sitios eliminados

Nuestras bases de datos cuentan con una combinación de copias de seguridad completas y copias de seguridad incrementales que nos permiten elegir cualquier «punto en el tiempo» específico para recuperar nuestros almacenes de datos dentro del período de retención de copias de seguridad (30 días). Para la mayoría de clientes, durante este incidente, identificamos los principales almacenes de datos de nuestros productos y decidimos utilizar un punto de restauración de cinco minutos antes de la eliminación de los sitios como punto de sincronización seguro. Los almacenes de datos no principales se restauraron al mismo punto o repitiendo los eventos registrados. El uso de un punto de restauración fijo para los almacenes principales permitió que los datos fueran coherentes en todos los almacenes de datos.

Para 57 clientes para los que la restauración tuvo lugar al principio de nuestra respuesta ante incidentes, la falta de políticas consistentes y la recuperación manual de las instantáneas de copias de seguridad de la base de datos hizo que algunas bases de datos de Confluence e Insight se restauraran a un punto con una anterioridad de *más* de cinco minutos respecto a la eliminación del sitio. La inconsistencia se descubrió durante un proceso de auditoría posterior a la restauración. Desde entonces, hemos recuperado el resto de los datos, nos hemos puesto en contacto con los clientes afectados y les estamos ayudando a aplicar cambios para continuar restaurando sus datos.

En resumen:

- Cumplimos nuestro objetivo de punto de recuperación (RPO) de una hora durante este incidente.

- La pérdida de datos a causa del incidente tiene un límite de cinco minutos antes de la eliminación del sitio.
- A un pequeño número de clientes se les restauraron las bases de datos de Confluence o Insight a un punto con una anterioridad de más de cinco minutos respecto a la eliminación del sitio; sin embargo, podemos recuperar los datos y actualmente estamos trabajando con los clientes para restaurarlos.

Comunicación de incidentes

Para nosotros, la comunicación de incidentes abarca puntos de contacto con clientes, socios, medios de comunicación, analistas del sector, inversores y la comunidad tecnológica en general.

Qué pasó

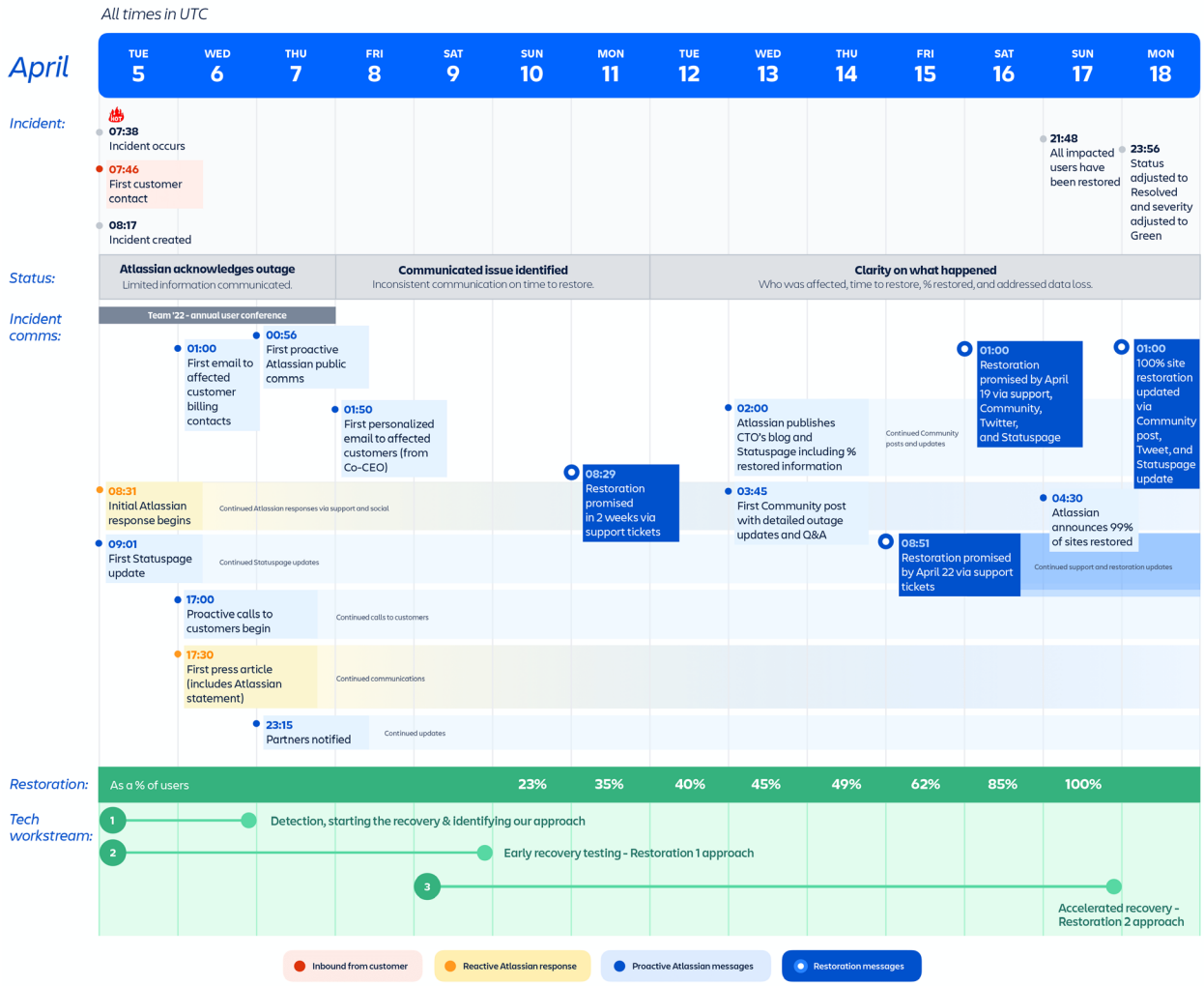


Figura 9: cronograma de los hitos clave de la comunicación de incidentes.

Marca de tiempo: días 1-3 (del 5 al 7 de abril)

Respuesta temprana

El primer ticket de asistencia se creó el 5 de abril a las 7:46 UTC y el soporte de Atlassian respondió admitiendo el incidente a las 8:31 UTC. A las 9:03 UTC, se publicó la primera actualización de Statuspage para informar a los clientes de que estábamos investigando el incidente. A las 11:13 UTC, confirmamos a través de Statuspage que habíamos identificado la causa principal y que estábamos trabajando en una solución. Antes de la 1:00 UTC del 6 de abril, las comunicaciones iniciales de tickets de los clientes indicaban que la interrupción se debía a un script de mantenimiento y que esperábamos una pérdida de datos mínima. Atlassian respondió a las consultas de los medios con un comunicado el 6 de abril a las 17:30 UTC. Atlassian tuiteó su primer mensaje externo amplio en el que admitía el incidente el 7 de abril a las 00:56 UTC.

Marca de tiempo: días 4-7 (del 8 al 11 de abril)

Comienza un alcance más amplio y personalizado

El 8 de abril a las 1:50 UTC, Atlassian envió por correo electrónico a los clientes afectados una disculpa del cofundador y codirector ejecutivo, Scott Farquhar. En los días siguientes, trabajamos para restaurar la información de contacto eliminada y crear tickets de asistencia para todos los sitios afectados que aún no habían presentado ninguno. Luego, nuestro equipo de soporte continuó enviando actualizaciones periódicas sobre los esfuerzos de restauración a través de los tickets de asistencia asociados con cada sitio afectado.

Marca de tiempo: días 8-14 (12 al 18 de abril)

Mayor claridad y restauración completa

El 12 de abril, Atlassian publicó una actualización del director de tecnología (CTO), Sri Viswanath, en la que se proporcionaban más detalles técnicos sobre qué había sucedido, quién se había visto afectado, si se había producido una pérdida de datos y cómo avanzaba el proceso de restauración, a la vez que se informaba de que la restauración completa de todos los sitios podría tomar hasta dos semanas. El blog se acompañó con otro comunicado de prensa atribuido a Sri. También hicimos referencia al blog de Sri en nuestra primera publicación proactiva de la Comunidad de Atlassian por parte del responsable de ingeniería, Stephen Deasy, que posteriormente se convirtió en el lugar dedicado a actualizaciones adicionales y preguntas y respuestas con el público en general. La restauración completa de todos los sitios de clientes afectados se anunció el 18 de abril en una actualización de esta publicación.

¿Por qué no respondimos antes públicamente?

1. Priorizamos la comunicación directa con los clientes afectados a través de Statuspage, correo electrónico, tickets de asistencia e interacciones individuales. Sin embargo, no pudimos ponernos en contacto con muchos clientes porque perdimos sus datos de contacto cuando se eliminaron sus sitios. Deberíamos haber implementado comunicaciones más amplias mucho antes para informar a los clientes y usuarios finales afectados sobre nuestro cronograma de respuesta y resolución de incidentes.

2. Si bien supimos de inmediato qué había causado el incidente, la complejidad arquitectónica y las circunstancias únicas de este incidente ralentizaron nuestra capacidad de analizar rápidamente y estimar con precisión el tiempo de resolución. En lugar de esperar a tener una imagen completa, deberíamos haber sido transparentes sobre lo que sí sabíamos y lo que no sabíamos. Proporcionar estimaciones generales de restauración (aunque solo fueran direccionales) y dejar claro cuándo esperábamos tener un panorama más completo habría permitido que nuestros clientes gestionaran mejor sus planes en relación con el incidente. Esto es particularmente cierto para los administradores de sistemas y los contactos técnicos, que están en primera línea en la gestión de las partes interesadas y los usuarios dentro de sus organizaciones.

Experiencia de asistencia y contacto con el cliente

Como se ha mencionado anteriormente, el mismo script que eliminó sitios de clientes también eliminó identificadores clave de los clientes y la información de contacto (p. ej. La URL de Cloud, los contactos del administrador de sistemas del sitio) de nuestros entornos de producción. Esto es importante porque nuestros sistemas principales (p. ej. soporte, licencias, facturación) aprovechan la existencia de una URL de Cloud y los contactos del administrador de sistemas del sitio como identificadores principales con fines de seguridad, enrutamiento y priorización. Cuando perdimos estos identificadores, perdimos inicialmente nuestra capacidad de identificar e interactuar sistemáticamente con los clientes.

¿Cómo se vio afectada la asistencia para nuestros clientes?

En primer lugar, la mayoría de los clientes afectados no podían comunicarse con nuestro equipo de soporte a través del formulario de contacto en línea normal. Este formulario está diseñado para pedir al usuario que inicie sesión con su ID de Atlassian y que proporcione una URL de Cloud válida. Sin una URL válida, el usuario no puede enviar un ticket de asistencia técnica. En circunstancias normales, esta verificación es intencional para garantizar la seguridad del sitio y la clasificación de los tickets. Sin embargo, este requisito creó un resultado no deseado para los clientes afectados por esta interrupción, ya que se les impidió enviar un ticket de asistencia del sitio de alta prioridad.

En segundo lugar, la eliminación de los datos del administrador de sistemas del sitio causada por el incidente creó una brecha en nuestra capacidad de interactuar proactivamente con los clientes afectados. En los primeros días del incidente, enviamos comunicaciones proactivas a los contactos técnicos y de facturación de los clientes afectados registrados en Atlassian. Sin embargo, nos percatamos rápidamente de que muchos contactos técnicos y de facturación de los clientes afectados estaban desactualizados. Sin la información del administrador de sistemas de cada sitio, no teníamos una lista completa de contactos activos y aprobados con los que ponernos en contacto.

¿Cómo respondimos?

Nuestros equipos de soporte tenían tres prioridades de igual importancia para acelerar la restauración del sitio y reparar la avería en nuestros canales de comunicación en los primeros días del incidente.

Primero, obtener una lista fiable de contactos de clientes validados. Mientras nuestros equipos de ingeniería trabajaban para restaurar los sitios de clientes, nuestros equipos de atención al cliente se centraron en restaurar la información de contacto validada. Utilizamos todos los mecanismos de los que disponíamos (sistemas de facturación, tickets de asistencia previos, otras copias de seguridad de los usuarios, contacto directo con los clientes, etc.) para reconstruir nuestra lista de contactos. Nuestro objetivo era tener un ticket de asistencia relacionado con incidentes para cada sitio afectado a fin de agilizar el contacto directo y los tiempos de respuesta.

En segundo lugar, restablecer los flujos de trabajo, las colas y los SLA específicos de este incidente. La eliminación del ID de Cloud y la incapacidad de autentificar correctamente a los usuarios también afectaron a nuestra capacidad de procesar tickets de asistencia relacionados con incidentes a través de nuestros sistemas normales. Los tickets no aparecían correctamente en las colas y paneles de prioridad y escalación relevantes. Creamos rápidamente un equipo multifuncional (soporte, producto, TI) para diseñar y agregar lógica, SLA, estados de flujo de trabajo y paneles adicionales.

Como esto tenía que hacerse dentro de nuestro sistema de producción, nos llevó varios días desarrollarlo, probarlo e implementarlo por completo.

En tercer lugar, escalar masivamente las validaciones manuales para acelerar las restauraciones del sitio. A medida que el equipo de ingeniería avanzaba en las restauraciones iniciales, quedó claro que se necesitaría la capacidad de nuestros equipos de soporte globales para ayudar a acelerar la recuperación de sitios mediante pruebas manuales y comprobaciones de validación. Este proceso de validación se convirtió en una vía crucial para que nuestros clientes pudieran restaurar sus sitios en cuanto nuestro equipo de ingeniería acelerara las restauraciones de datos. Tuvimos que crear un flujo independiente de procedimientos operativos estándar (SOP), flujos de trabajo, transferencias y listas de personal para movilizar a más de 450 ingenieros de soporte para realizar comprobaciones de validación, con turnos que proporcionaban cobertura ininterrumpida, para acelerar las restauraciones y entregárselas a los clientes.

Incluso con estas prioridades clave bien establecidas al final de la primera semana, nuestra capacidad de proporcionar actualizaciones *significativas* era limitada, ya que faltaba claridad en torno a los plazos de resolución de incidentes debido a la complejidad de los procesos de restauración. Deberíamos haber admitido antes nuestra incertidumbre a la hora de proporcionar una fecha de restauración del sitio y habernos puesto antes a disposición de los clientes para hablarlo en persona de modo que pudieran hacer planes en consecuencia.

¿Cómo mejoraremos?

Hemos bloqueado inmediatamente las eliminaciones masivas de sitios hasta que se puedan realizar los cambios adecuados.

A medida que vamos dejando atrás este incidente y revaluamos nuestros procesos internos, queremos reconocer que las personas no causan incidentes. Por el contrario, los sistemas permiten que se cometan errores. Esta sección resume los factores que contribuyeron a este incidente. También abordamos nuestros planes para acelerar la forma en que solucionaremos estas deficiencias y problemas.

Aprendizaje 1: los «borrados suaves» deben ser universales en todos los sistemas

En general, la eliminación de este tipo debe estar prohibida o tener múltiples capas de protección para evitar errores. La principal mejora que estamos realizando es evitar globalmente la eliminación de los datos y metadatos de los clientes que no hayan pasado por un proceso de borrado suave.

a) La eliminación de datos solo debe realizarse como un borrado suave
Debe prohibirse la eliminación de todo un sitio y el borrado suave debe requerir protecciones de varios niveles para evitar errores. Implementaremos una política de «borrado suave», que evitará que los scripts o sistemas externos eliminen los datos de los clientes en un entorno de producción. Nuestra política de «borrado suave» permitirá una retención de datos suficiente para que la recuperación de datos se pueda ejecutar de forma rápida y segura. Los datos solo se eliminarán del entorno de producción después de que haya expirado el período de retención.

Acciones:
- **Implementar un «borrado suave» en los flujos de trabajo de aprovisionamiento y en todos los almacenes de datos relevantes.** Además, el equipo de Tenant Platform verificará que las eliminaciones de datos solo puedan ocurrir después de las desactivaciones, así como otras medidas de seguridad en este espacio. A largo plazo, Tenant Platform asumirá un papel de liderazgo para seguir desarrollando una correcta gestión del estado de los datos de los inquilinos.

b) El borrado suave debe tener un proceso de revisión estandarizado y verificado
Las acciones de borrado suave son operaciones de alto riesgo. Como tal, debemos tener procesos de revisión estandarizados o automatizados que incluyan procedimientos de prueba y de reversiones definidos para abordar estas operaciones.

Acciones:
- **Implementación por etapas obligatoria de cualquier acción de borrado suave:** todas las operaciones nuevas que requieran algún tipo de eliminación se probarán primero en nuestros propios sitios para validar nuestro enfoque y verificar la automatización. Una vez que hayamos completado esa validación, haremos que los clientes pasen progresivamente por el mismo proceso y continuaremos realizando pruebas para detectar irregularidades antes de aplicar la automatización a toda la base de usuarios seleccionada.
- **Las acciones de borrado suave deben tener un plan de reversión probado:** cualquier actividad de borrado suave de datos debe probar la restauración de los datos eliminados antes de ejecutarse en producción y tener un plan de reversión probado.

Aprendizaje 2: como parte del programa de DR, automatizar la restauración de eventos de eliminación de múltiples sitios y productos para un conjunto más grande de clientes

La gestión de datos de Atlassian proporciona una descripción detallada de nuestros procesos de gestión de datos. Para conseguir una alta disponibilidad, aprovisionamos y mantenemos una réplica en espera sincrónica en varias zonas de disponibilidad (AZ) de AWS. La conmutación por error de AZ está automatizada y suele tardar entre 60 y 120 segundos, y gestionamos regularmente las interrupciones del servicio de centros de datos y otras interrupciones habituales sin que esto afecte a los clientes.

También mantenemos copias de seguridad inmutables que están diseñadas para resistir eventos de corrupción de datos, lo que permite la restauración a un punto anterior en el tiempo. Las copias de seguridad se conservan durante 30 días y Atlassian prueba y audita continuamente la restauración de copias de seguridad de almacenamiento. Y, si es necesario, podemos restaurar a todos los clientes a la vez en un entorno nuevo.

Con estas copias de seguridad, revertimos habitualmente clientes individuales o un pequeño conjunto de clientes que eliminan accidentalmente sus datos. Sin embargo, la eliminación a nivel de sitio no tenía runbooks que pudieran automatizarse rápidamente para la escala de este evento, que requería que las herramientas y la automatización en todos los productos y servicios tuvieran lugar de manera coordinada.

Lo que aún no hemos automatizado es la restauración de un gran subconjunto de clientes en el entorno que usamos actualmente sin afectar al resto de los clientes.

Dentro de nuestro entorno en la nube, cada almacén de datos contiene datos de varios clientes. Dado que los datos eliminados en este incidente eran solo una parte de los almacenes de datos que otros clientes siguen utilizando, tenemos que extraer y restaurar manualmente partes individuales de nuestras copias de seguridad. La recuperación del sitio de cada cliente es un proceso largo y complejo que requiere validación interna y verificación final del cliente una vez que se restaura el sitio.

Acciones:
Acelerar las restauraciones de múltiples productos y sitios para un conjunto más grande de clientes: el programa de DR cumple con nuestros estándares actuales de RPO de una hora. Aprovecharemos la automatización y los aprendizajes de este incidente para acelerar el programa de DR a fin de cumplir con el RTO definido en nuestra política para esta escala de incidentes.

Automatizar y agregar la verificación de este caso a las pruebas de DR: realizaremos regularmente ejercicios de DR que impliquen la restauración de todos los productos para un gran conjunto de sitios. Estas pruebas de DR verificarán que los runbooks estén actualizados a medida que nuestra arquitectura evolucione y se encuentren nuevos casos extremos. Mejoraremos continuamente nuestro enfoque de restauración, automatizaremos más partes del proceso de restauración y reduciremos el tiempo de recuperación.

Aprendizaje 3: mejorar el proceso de gestión de incidentes para eventos a gran escala

Nuestro programa de gestión de incidentes es adecuado para gestionar los incidentes graves y menores que se han producido a lo largo de los años. Simulamos con frecuencia la respuesta ante incidentes para eventos de menor escala y menor duración, que normalmente involucran a menos personas y equipos.

Sin embargo, en su punto álgido, este incidente hizo que cientos de ingenieros y empleados de atención al cliente tuvieran que trabajar simultáneamente para restaurar sitios de clientes. Nuestro programa y equipos de gestión de incidentes no se diseñaron para gestionar la profundidad, la amplitud y la duración de este tipo de incidente (consulta la *figura 10* a continuación).

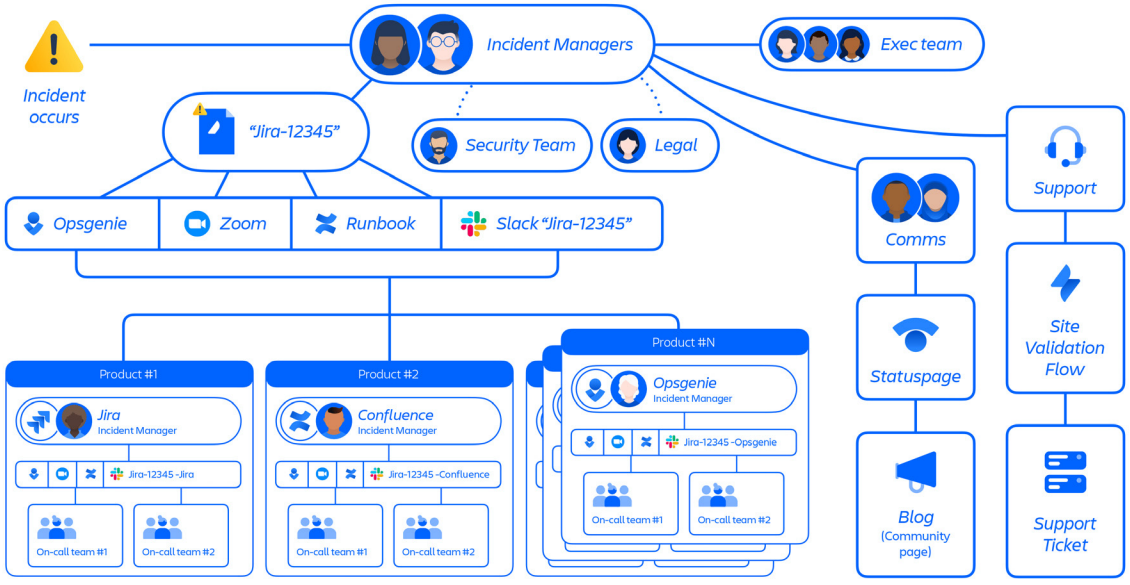


Figura 10: descripción general del proceso de gestión de incidentes a gran escala.

Nuestro proceso de gestión de incidentes a gran escala estará mejor definido y se practicará con frecuencia

Tenemos manuales de estrategias para incidentes a nivel de producto, pero no para eventos de esta escala, con cientos de personas trabajando simultáneamente en toda la empresa. En las herramientas de gestión de incidentes, tenemos una automatización que crea flujos de comunicación como Slack, Zoom y Confluence doc, pero no crea los subflujos que se requieren en incidentes a gran escala para aislar los flujos de restauración.

Acciones:
Definir un manual de estrategias y herramientas para incidentes a gran escala y realizar ejercicios simulados: definir y documentar los tipos de incidentes que pueden considerarse de gran escala y que requieren este nivel de respuesta. Describir los pasos clave de coordinación y generar herramientas para ayudar a los gestores de incidentes y otras funciones empresariales a optimizar la respuesta e iniciar la recuperación. Los gestores de incidentes y sus equipos realizarán simulaciones y formaciones, y perfeccionarán sus herramientas y documentos con regularidad para garantizar una mejora continua.

Aprendizaje 4: mejorar nuestros procesos de comunicación

a) Eliminamos identificadores esenciales de clientes, lo que tuvo un impacto en las comunicaciones y las acciones de los afectados
El mismo script que eliminó sitios de clientes también eliminó identificadores clave de los clientes (p. ej. la URL de Cloud, los contactos del administrador de sistemas del sitio) de nuestros entornos de producción. Como resultado, (1) los clientes no podían presentar tickets de asistencia técnica a través de nuestro canal de soporte normal; (2) nos llevó días obtener una lista fiable de los contactos clave de los clientes (como los administradores de sistemas del sitio) afectados por la interrupción de la participación proactiva; y (3) en un primer momento, los flujos de trabajo de asistencia, SLA, paneles y procesos de escalación no funcionaban correctamente debido a la naturaleza única del incidente.

Durante la interrupción, las escalaciones de clientes también llegaron a través de varios canales (correo electrónico, llamadas telefónicas, tickets de directores ejecutivos, tickets de asistencia y LinkedIn y otros canales sociales). Las herramientas y procesos dispares en nuestros equipos de atención al cliente ralentizaron nuestra respuesta y dificultaron el seguimiento y la generación de informes integrales de estas escalaciones.

b) No teníamos un manual de estrategias de comunicación de incidentes lo suficientemente completo como para hacer frente a este nivel de complejidad

No teníamos un manual de estrategias de comunicación de incidentes que describiera tanto los principios como las funciones y las responsabilidades necesarios para movilizar un equipo de comunicación de incidentes unificado y multifuncional con la suficiente rapidez. No reconocimos el incidente de forma rápida y coherente a través de los diferentes canales, especialmente en las redes sociales. El enfoque correcto habría sido ofrecer una comunicación pública más amplia en relación con la interrupción, además de repetir el mensaje fundamental de que no hubo pérdida de datos y que la interrupción no fue provocada por un ciberataque.

Acciones:

- **Mejorar la copia de seguridad de los contactos clave:** realizar una copia de seguridad de la información de contacto de la cuenta autorizada fuera de la instancia de producto.
- **Herramientas de soporte actualizadas:** crear mecanismos para que los clientes que no tengan una URL de sitio o un ID de Atlassian válidos se pongan en contacto directamente con nuestro equipo de asistencia técnica.
- **Sistema y procesos de escalación de clientes:** invertir en un sistema de escalación unificado, basado en cuentas y en flujos de trabajo que permitan almacenar múltiples objetos de trabajo (tickets, tareas, etc.) en un único objeto de cuenta de cliente, con lo que mejorará la coordinación y la visibilidad de todos nuestros equipos de atención al cliente.
- **Agilizar la cobertura ininterrumpida de gestión de escalaciones:** ejecutar planes de expansión de la presencia global de la función de gestión de escalaciones para garantizar una cobertura uniforme las 24 horas del día, los 7 días de la semana, con personal designado ubicado en las principales regiones geográficas junto a funciones de apoyo que contribuyan con conocimientos expertos en productos y ventas y el liderazgo requeridos.
- **Actualizar nuestro manual de estrategias de comunicación de incidentes con lo que vayamos aprendiendo y revisarlo con regularidad:** revisar el manual de estrategias para definir funciones y líneas de comunicación claras internamente. Utilizar el marco DACI para incidentes y tener respaldos las 24 horas del día, los 7 días de la semana, para todas las funciones en caso de enfermedad, vacaciones u otros eventos imprevistos. Llevar a cabo una auditoría trimestral para verificar la disponibilidad en todo momento.

Acciones (continuación)
Seguir la plantilla de comunicación de incidentes en todas las comunicaciones: exponer lo que sucedió, quién se vio afectado, el cronograma de la restauración, los porcentajes de restauración del sitio, la pérdida de datos esperada y los niveles de confianza asociados, y proporcionar una orientación clara sobre cómo contactar con el soporte.

Observaciones finales

Aunque se haya resuelto la interrupción y los datos de los clientes se hayan recuperado por completo, nuestro trabajo continúa. En este momento, estamos implementando los cambios descritos anteriormente para mejorar nuestros procesos, aumentar nuestra resiliencia y evitar que una situación como esta vuelva a ocurrir.

Atlassian es una organización en constante aprendizaje y, sin duda, nuestros equipos han aprendido muchas lecciones difíciles de esta experiencia. Estamos poniendo en práctica todo lo aprendido para hacer cambios duraderos en nuestra empresa. En última instancia, nos fortaleceremos y te proporcionaremos un mejor servicio gracias a esta experiencia.

Esperamos que lo que hemos aprendido de este incidente sea útil para otros equipos que trabajan diligentemente para brindar servicios fiables a sus clientes.

Por último, quiero dar las gracias a quienes leen esto y aprenden con nosotros, y a quienes forman parte de nuestra amplia comunidad y equipo de Atlassian.

-Sri Viswanath, director de tecnología