ATLASSIAN

Análise pós-incidente

Interrupção de abril de 2022

Esta tradução é disponibilizada apenas por conveniência. No caso de qualquer ambiguidade ou conflito entre as traduções, a versão original em inglês prevalece.

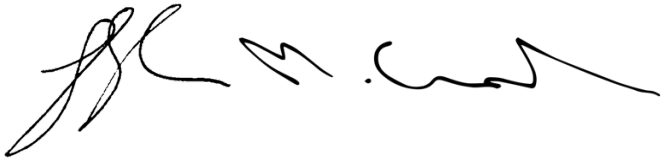
Carta dos cofundadores e co-CEOs

A gente quer reconhecer a interrupção que suspendeu o serviço para os clientes no início deste mês. A gente entende que os produtos da Atlassian são essenciais para a sua empresa e leva essa responsabilidade a sério. O dinheiro para com a gente. Ponto final. Para os clientes afetados, a gente está trabalhando para recuperar sua confiança.

Na Atlassian, um dos principais valores é “Uma empresa aberta, sem papo-furado”. A gente dá vida a esse valor, em parte, discutindo abertamente os incidentes e usando-os como oportunidades de aprendizado. A gente está publicando esta análise pós-incidente para os clientes, a comunidade da Atlassian e a comunidade técnica mais ampla. A Atlassian tem orgulho de ter um processo de gerenciamento de incidentes que enfatiza que uma cultura que não aponta culpados e o foco na identificação de maneiras de melhorar os sistemas e processos técnicos são essenciais para oferecer serviços confiáveis e de alta escala. Embora a gente faça o máximo para evitar qualquer tipo de incidente, a gente também adota a ideia de que os incidentes são uma maneira poderosa de melhorar.

Fique tranquilo, a plataforma de nuvem da Atlassian permite que a gente atenda às diversas necessidades dos mais de 200.000 clientes de nuvem de todos os tamanhos e em todos os setores. Antes deste incidente, a nuvem disponibilizava consistentemente 99,9% de tempo de atividade e superava os SLAs de tempo de atividade. A gente fez investimentos de longo prazo na plataforma e em vários recursos de plataforma centralizada, com uma infraestrutura escalável e uma cadência constante de aprimoramentos de segurança.

Aos clientes e parceiros, a gente agradece pela confiança e pela parceria contínuas. A gente espera que os detalhes e as ações descritas neste documento mostrem que a Atlassian vai continuar a oferecer uma plataforma de nuvem de classe mundial e um poderoso portfólio de produtos para atender às necessidades de cada equipe.



-Scott e Mike

Resumo executivo

Na terça-feira, 5 de abril de 2022, a partir das 7:38 UTC, 775 clientes da Atlassian perderam o acesso aos seus produtos da Atlassian. A interrupção durou até 14 dias para um subconjunto desses clientes, com o primeiro conjunto de clientes tendo os sites restaurados em 8 de abril e todos os sites de clientes sendo restaurados progressivamente até 18 de abril.

Essa interrupção não foi resultado de um ataque cibernético e não houve acesso não autorizado aos dados dos clientes. A Atlassian tem um programa abrangente de gerenciamento de dados com SLAs publicados e um histórico de superação desses SLAs.

Embora esse tenha sido um incidente grave, nenhum cliente perdeu mais de cinco minutos de dados. Além disso, mais de 99,6% dos clientes e usuários continuaram a usar os produtos da nuvem da Atlassian sem qualquer interrupção durante as atividades de restauração.

> Ao longo deste documento, a gente fala em "afetados" e "impactados" para se referir aos clientes que tiveram os sites excluídos como parte deste incidente. Esta análise pós-incidente mostra os detalhes exatos do incidente, apresenta as etapas pelas quais a gente passou para a recuperação e descreve como a gente vai evitar que situações como essa aconteçam no futuro. A gente disponibiliza um resumo geral do incidente nesta seção, com mais detalhes no restante do documento.

O que aconteceu?

Em 2021, a gente concluiu a aquisição e a integração de um aplicativo independente da Atlassian para o Jira Service Management e o Jira Software chamado "Insight - Asset Management". A funcionalidade desse aplicativo independente era então nativa no Jira Service Management e não estava mais disponível para o Jira Software. Por causa disso, a gente precisava excluir o aplicativo legado independente nos sites dos clientes que o tinham instalado. As equipes de engenharia usaram um script e um processo existentes para excluir instâncias desse aplicativo independente, mas havia dois problemas:

- **Falha de comunicação.** Primeiro, havia uma falha de comunicação entre a equipe que solicitou a exclusão e a equipe que executou a exclusão. Em vez de oferecer

os IDs do *aplicativo em questão* marcado para exclusão, a equipe disponibilizou os IDs de *todo o site de nuvem* em que os aplicativos deveriam ser excluídos.

- **Avisos insuficientes do sistema.** A API usada para executar a exclusão aceitou identificadores de site e aplicativo e presumiu que a entrada estava correta – ou seja, se um ID de site era passado, um site era excluído; se um ID de aplicativo era passado, um aplicativo era excluído. Não houve sinal de aviso para confirmar o tipo de exclusão (site ou aplicativo) que estava sendo solicitado.

O script que foi executado seguiu o processo padrão de revisão por pares da Atlassian, com foco em qual ponto de extremidade estava sendo chamado e como. Ele não cruzou os IDs de site em nuvem informados para verificar se eles se referiam ao Insight App ou ao site inteiro, e o problema foi que o script continha o ID de todo o site de um cliente. O resultado foi uma exclusão imediata de 883 sites (representando 775 clientes) entre 07:38 UTC e 08:01 UTC na terça-feira, 5 de abril de 2022. *Veja "O que aconteceu"*

Como a gente respondeu?

Assim que o incidente foi confirmado em 5 de abril às 08:17 UTC, a gente acionou o processo de gerenciamento de incidentes graves que a gente tem e formou uma equipe multifuncional de gerenciamento de incidentes. A equipe global de resposta a incidentes trabalhou 24 horas por dia, 7 dias por semana durante o incidente, até que todos os sites fossem restaurados, validados e devolvidos aos clientes. Além disso, os líderes de gerenciamento de incidentes faziam reuniões a cada três horas para coordenar os fluxos de trabalho.

Logo no início, a gente percebeu que havia vários desafios para restaurar centenas de clientes com vários produtos ao mesmo tempo.

No início do incidente, a gente sabia exatamente quais sites tinham sido afetados e teve como prioridade estabelecer comunicação com o proprietário aprovado de cada site afetado para informá-lo sobre a interrupção.

No entanto, algumas informações de contato do cliente foram excluídas, então os clientes não podiam registrar tickets de suporte como fariam normalmente, e a gente não tinha acesso imediato aos principais contatos dos clientes. *Para mais informações, consulte "Visão geral de alto nível dos fluxos de trabalho de recuperação"*

O que a gente está fazendo para evitar situações como essa no futuro?

A gente tomou uma série de ações imediatas e estabeleceu o compromisso de fazer mudanças para evitar essa situação no futuro. Aqui estão quatro áreas específicas nas quais a gente fez ou vai fazer mudanças significativas:

1. **Estabelecer "exclusões temporárias" universais em todos os sistemas**. No geral, uma exclusão desse tipo deve ser proibida ou ter várias camadas de proteção para evitar erros, incluindo implementação em etapas e um plano testado de rollback para "exclusões temporárias". A gente vai impedir globalmente a exclusão de dados e metadados dos clientes que não tenham passado por um processo de exclusão temporária.

2. **Acelerar o programa de recuperação de desastres (DR) para automatizar a restauração de eventos de exclusão de vários sites e vários produtos para um conjunto maior de clientes.** A gente vai aproveitar a automação e os aprendizados deste incidente para acelerar o programa de DR e atender ao RTO (Recovery Time Objective, objetivo de tempo de recuperação), conforme definido na política que a gente tem para essa escala de incidentes. A gente vai executar regularmente exercícios de DR que envolvam a restauração de todos os produtos para um grande conjunto de sites.

3. **Revisar o processo de gerenciamento de incidentes para incidentes de grande escala**. A gente vai melhorar o procedimento operacional padrão para incidentes de grande escala e praticar esse procedimento com simulações dessa escala de incidente. A gente também vai atualizar o treinamento e as ferramentas para lidar com o grande número de equipes trabalhando em paralelo.

4. **Criar um esquema tático de comunicação de incidentes em grande escala**. A gente vai reconhecer os incidentes com antecedência, por meio de vários canais. A gente vai divulgar comunicações públicas sobre incidentes em poucas horas. Para alcançar melhor os clientes afetados, a gente vai melhorar o backup dos principais contatos e ajustar as ferramentas de suporte para permitir que clientes sem URL válida ou ID da Atlassian entrem em contato direto com a equipe de suporte técnico.

A lista completa de itens de ação está detalhada na análise completa pós-incidente abaixo. *Veja "Como a gente vai melhorar"*

Sumário

Visão geral da arquitetura de nuvem da Atlassian Página 7
- Arquitetura de hospedagem em nuvem da Atlassian
- Arquitetura de serviços distribuídos
- Arquitetura de vários locatários
- Provisionamento de locatário e ciclo de vida
- Programa de recuperação de desastres
    - Resiliência
    - Capacidade de restauração de armazenamento do serviço
    - Capacidade de restauração automatizada em vários sites e vários produtos

O que aconteceu, linha do tempo e recuperação Página 13
- O que aconteceu
- Como foi a coordenação
- Linha do tempo do incidente
- Visão geral de alto nível dos fluxos de trabalho de recuperação
    - Fluxo de trabalho 1: detecção, início da recuperação e identificação da abordagem
    - Fluxo de trabalho 2: recuperação inicial e abordagem Restauração 1
    - Fluxo de trabalho 3: recuperação acelerada e abordagem Restauração 2
    - Perda mínima de dados após a restauração de sites excluídos

Comunicações de incidentes Página 21
- O que aconteceu

Experiência de suporte e contato com o cliente Página 23
- Como o suporte aos clientes foi afetado?
- Como a gente respondeu?

Como a gente vai melhorar? Página 25
- Aprendizado 1: as “exclusões temporárias” devem ser universais em todos os sistemas
- Aprendizado 2: como parte do programa de recuperação de desastres, automatizar a restauração para eventos de exclusão de vários sites e vários produtos para um conjunto maior de clientes
- Aprendizado 3: melhorar o processo de gerenciamento de incidentes para eventos de grande escala
- Aprendizado 4: melhorar os processos de comunicação

Considerações finais Página 31

Visão geral da arquitetura de nuvem da Atlassian

Para entender os fatores que contribuíram para o incidente, conforme discutido ao longo deste documento, é útil entender primeiro a arquitetura de implementação dos produtos, serviços e infraestrutura da Atlassian.

Arquitetura de hospedagem em nuvem da Atlassian

A Atlassian usa a Amazon Web Services (AWS) como provedor de serviços de nuvem e suas instalações de data center com alta disponibilidade em várias regiões do mundo. Cada região da AWS é uma localização geográfica separada com grupos de data centers múltiplos, isolados e fisicamente separados, conhecidos como zonas de disponibilidade (AZs).

A gente aproveita os serviços de computação, armazenamento, rede e dados da AWS para criar produtos e componentes de plataforma, o que permite que a gente utilize recursos de redundância oferecidos pela AWS, como regiões e zonas de disponibilidade.

Arquitetura de serviços distribuídos

Com essa arquitetura da AWS, a gente hospeda diversos produtos e serviços de plataforma que são usados nas soluções da Atlassian. Entre eles estão recursos de plataforma que são compartilhados e consumidos em vários produtos da Atlassian, como Mídia, Identidade, Comércio, experiências como o Editor, bem como recursos específicos do produto, como o Jira Issue Service e o Confluence Analytics.

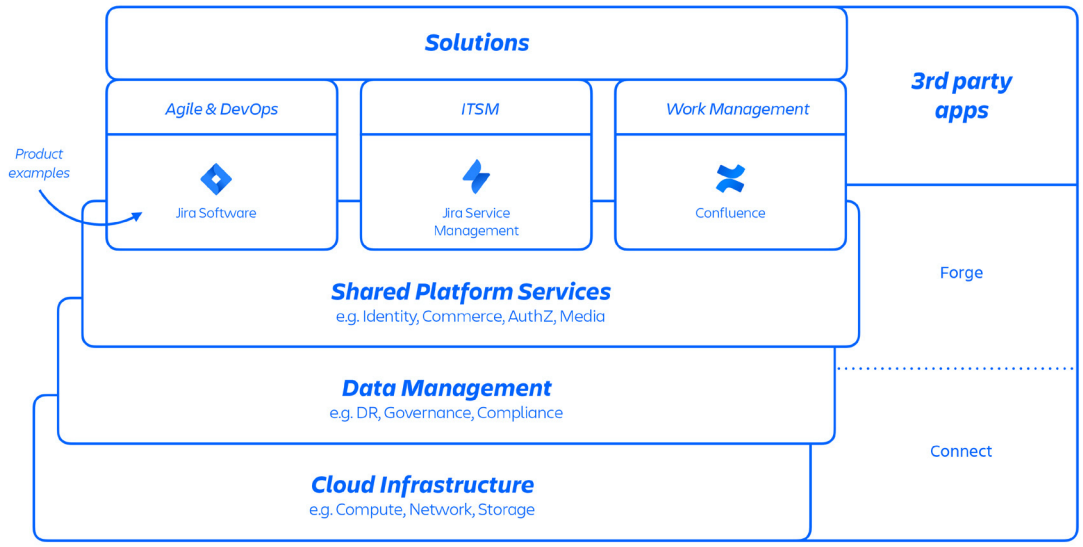


Figura 1: Arquitetura da plataforma da Atlassian.

Os desenvolvedores da Atlassian provisionam esses serviços por meio de uma plataforma como serviço (PaaS) desenvolvida internamente, chamada Micros, que orquestra automaticamente a implementação de serviços, infraestrutura, armazenamentos de dados compartilhados e seus recursos de gerenciamento, incluindo requisitos de controle de segurança e conformidade (veja a *Figura 1* acima). Normalmente, um produto da Atlassian consiste em vários serviços "conteinerizados" que são implementados na AWS usando o Micros. Os produtos da Atlassian usam recursos centrais de plataforma (veja a *Figura 2* abaixo) que variam de encaminhamento de solicitações a armazenamentos de objetos binários, autenticação/autorização, conteúdo gerado pelo usuário (UGC) transacional e armazenamentos de relacionamentos de entidades, data lakes, registro comum, rastreamento de solicitações, observabilidade e serviços de análise. Esses microsserviços são criados usando pilhas técnicas aprovadas padronizadas no nível da plataforma:

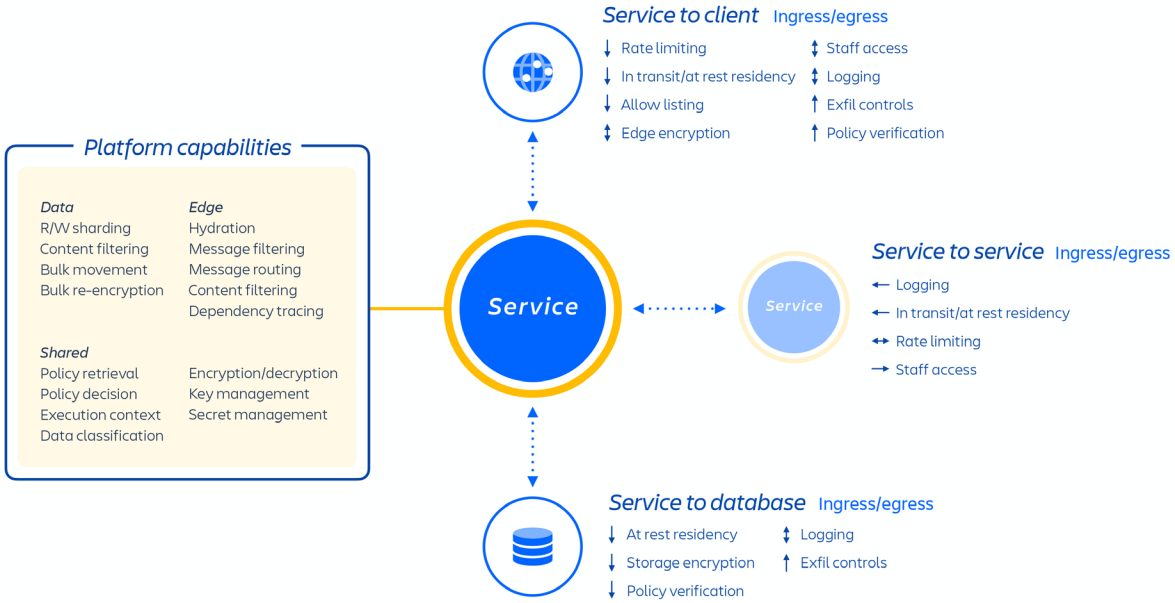


Figura 2: Visão geral dos microsserviços da Atlassian.

Arquitetura de vários locatários

Além da infraestrutura de nuvem, a gente criou e opera uma arquitetura de microsserviços de vários locatários junto com uma plataforma compartilhada que oferece suporte aos produtos da Atlassian. Em uma arquitetura de vários locatários, um único serviço atende a vários clientes, incluindo bancos de dados e instâncias de computação necessários para executar os produtos da nuvem. Cada fragmento (essencialmente um contêiner – veja a *Figura 3* abaixo) contém os dados de vários locatários, mas os dados de cada locatário são isolados e inacessíveis a outros locatários.

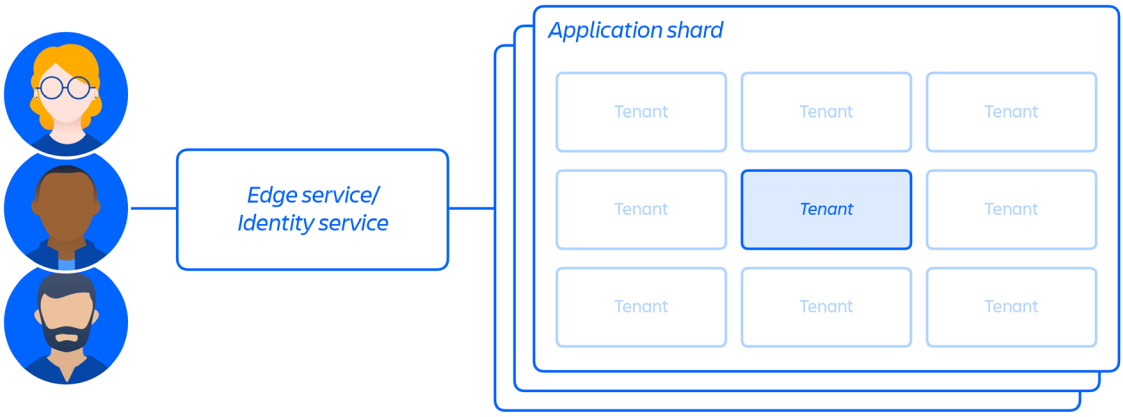


Figura 3: Como a gente armazena dados em uma arquitetura de vários locatários.

Provisionamento de locatário e ciclo de vida

Quando um novo cliente é provisionado, uma série de eventos aciona a orquestração de serviços distribuídos e o provisionamento de armazenamentos de dados. Esses eventos geralmente podem ser mapeados para uma das sete etapas do ciclo de vida:

1. Os sistemas de comércio são atualizados imediatamente com as informações de controle de acesso e os metadados mais recentes desse cliente; depois, um sistema de orquestração de provisionamento alinha o “estado dos recursos provisionados” com o estado da licença por meio de uma série de eventos de locatário e produto.

Eventos de locatário

Esses eventos afetam o locatário como um todo e podem ser:

- Criação: um locatário é criado e usado para novos sites
- Destruição: um locatário inteiro é excluído

Eventos de produtos

- Ativação: após a ativação de produtos licenciados ou aplicativos de terceiros
- Desativação: após a desativação de determinados produtos ou aplicativos
- Suspensão: após a suspensão de um determinado produto existente, desativando assim o acesso a um determinado site próprio
- Des-suspensão: após a des-suspensão de um determinado produto existente, permitindo assim o acesso a um site próprio

Atualização de licença: contém informações sobre o número de licenças para um determinado produto, bem como seu status (ativo/inativo)

2. Criação do site do cliente e ativação do conjunto correto de produtos para o cliente. O conceito de um site é o contêiner de vários produtos licenciados para um determinado cliente. (Por exemplo, o Confluence e o Jira Software para <nome-do-site>.atlassian.net.). Esse (veja a *Figura 4* abaixo) é um ponto importante a ser entendido no contexto deste relatório, já que o contêiner do site é o que foi excluído neste incidente e o conceito de um site é discutido ao longo do documento.

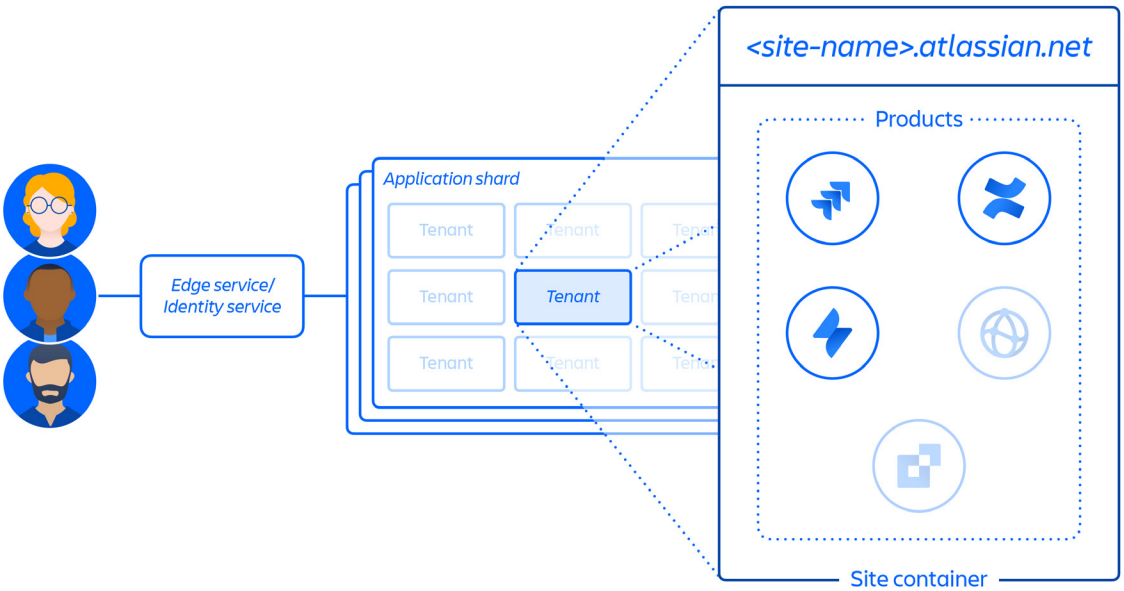


Figura 4: Visão geral do contêiner do site.

3. Provisionamento de produtos no site do cliente na região designada.

   Quando um produto é provisionado, ele vai ter a maior parte de seu conteúdo hospedado perto de onde os usuários o acessam. Para otimizar o desempenho do produto, a gente não limita a movimentação de dados quando eles são hospedados globalmente, podendo mover dados entre regiões conforme necessário.

   Para alguns produtos da Atlassian, a gente também oferece residência de dados. A residência de dados permite que os clientes escolham se os dados do produto são distribuídos globalmente ou mantidos no local em uma das localizações geográficas definidas que a gente oferece.

4. Criação e armazenamento dos principais metadados e configurações do(s) produto(s) e do site do cliente.

5. Criação e armazenamento dos dados de identidade do site e do(s) produto(s), como usuários, grupos, permissões, etc.

6. Provisionamento de bancos de dados de produtos em um site, por exemplo: família de produtos Jira, Confluence, Compass, Atlas.

7. Provisionamento dos aplicativos licenciados do(s) produto(s).

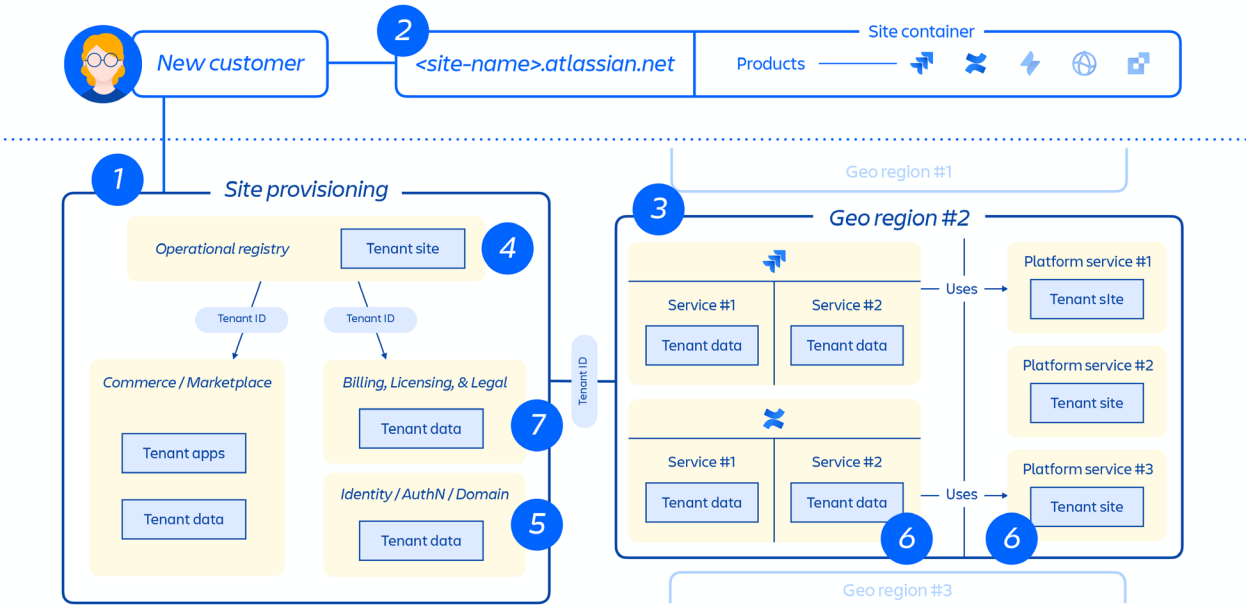


Figura 5: Visão geral de como o site do cliente é provisionado na arquitetura distribuída da Atlassian.

A *Figura 5* acima demonstra como o site de um cliente é implementado na arquitetura distribuída da Atlassian, não apenas em um único banco de dados ou armazenamento. São vários locais físicos e lógicos que armazenam metadados, dados de configuração, dados do produto, dados da plataforma e outras informações relacionadas ao site.

Programa de recuperação de desastres

O programa de recuperação de desastres (DR) engloba todos os esforços que a gente faz para oferecer resiliência contra falhas de infraestrutura e capacidade de restauração do armazenamento de serviços a partir de backups. Dois conceitos importantes para entender os programas de recuperação de desastres são:

- **Objetivo de tempo de recuperação (RTO):** com que rapidez os dados podem ser recuperados e devolvidos a um cliente durante um desastre?
- **Objetivo de ponto de recuperação (RPO):** os dados recuperados de um backup são recentes? Quanto dos dados vai ser perdidos desde o último backup?

Durante este incidente, a gente não atingiu o RTO, mas atingiu o RPO.

Resiliência

A gente se prepara para falhas no nível da infraestrutura; por exemplo, a perda de um banco de dados inteiro, de um serviço ou de zonas de disponibilidade da AWS. Essa preparação inclui a replicação de dados e serviços em várias zonas de disponibilidade e testes regulares de failover.

Capacidade de restauração de armazenamento do serviço

A gente também se prepara para se recuperar da corrupção de dados do armazenamento de serviços por riscos como ransomware, agentes mal-intencionados, defeitos de software e erros operacionais. Essa preparação inclui backups imutáveis e testes de restauração de backup de armazenamento de serviços. A gente pode restaurar qualquer armazenamento de dados individual para um ponto anterior no tempo.

Capacidade de restauração automatizada em vários sites e vários produtos

No momento do incidente, a gente não tinha a capacidade de selecionar um grande conjunto de sites de clientes e restaurar todos os seus produtos interconectados de backups para um ponto anterior no tempo.

A gente tem concentrado os recursos em infraestrutura, corrupção de dados, eventos de serviço único ou exclusões de um único site. No passado, a gente teve que abordar e testar esses tipos de falhas. A exclusão no nível do site não tinha runbooks que pudessem ser rapidamente automatizados para a escala desse evento, que exigia que ferramentas e automação em todos os produtos e serviços funcionassem de forma coordenada.

As seções a seguir vão abordar em mais detalhes essa complexidade e o que a gente está fazendo na Atlassian para evoluir e otimizar as habilidades para manter essa arquitetura em escala.

O que aconteceu, linha do tempo e recuperação

O que aconteceu

Em 2021, a gente concluiu a aquisição de um aplicativo independente da Atlassian para o Jira Service Management e o Jira Software chamado "Insight - Asset Management". A funcionalidade desse aplicativo independente era então nativa no Jira Service Management e não estava mais disponível para o Jira Software. Por causa disso, a gente precisava excluir o aplicativo legado independente nos sites dos clientes que o tinham instalado. As equipes de engenharia usaram um script e um processo existentes para excluir instâncias desse aplicativo independente.

No entanto, dois problemas críticos se seguiram:

- **Falha de comunicação.** Primeiro, havia uma falha de comunicação entre a equipe que solicitou a exclusão e a equipe que a executou. Em vez de oferecer os IDs do aplicativo em questão marcado para exclusão, a equipe disponibilizou os IDs de todo o site de nuvem em que os aplicativos deveriam ser excluídos.
- **Avisos insuficientes do sistema.** A API usada para executar a exclusão aceitou identificadores de site e aplicativo e presumiu que a entrada estava correta – ou seja, se um ID de site era passado, um site era excluído; se um ID de aplicativo era passado, um aplicativo era excluído. Não houve sinal de aviso para confirmar o tipo de exclusão (site ou aplicativo) que estava sendo solicitado.

O script que foi executado seguiu o processo padrão de revisão por pares da Atlassian, com foco em qual ponto de extremidade estava sendo chamado e como. Ele não cruzou os IDs de site em nuvem informados para verificar se eles se referiam ao aplicativo ou ao site inteiro. O script foi testado no staging de acordo com os processos de gerenciamento de alterações padrão; no entanto, ele não teria detectado que a entrada de IDs estava incorreta, pois os IDs não existiam no ambiente de staging.

Quando executado na produção, o script foi executado inicialmente para 30 sites. A primeira execução de produção foi bem-sucedida e excluiu o Insight App para esses 30 sites sem outros efeitos colaterais. No entanto, os IDs desses 30 sites foram obtidos antes do evento de falha de comunicação e incluíam os IDs corretos do Insight App.

O script para a execução de produção subsequente incluiu IDs de site no lugar de IDs do Insight App e foi executado para um conjunto de 883 sites. O script começou a ser executado

em 5 de abril às 07:38 UTC e foi concluído às 08:01 UTC. O script excluiu sites sequencialmente com base na lista de entrada, de modo que o site do primeiro cliente foi excluído logo após o script começar a ser executado às 07:38 UTC. O resultado foi uma exclusão imediata dos 883 sites, sem nenhum sinal de aviso para as equipes de engenharia.

Os seguintes produtos da Atlassian não estavam disponíveis para os clientes afetados: a família de produtos Jira, o Confluence, o Atlassian Access, o Opsgenie e o Statuspage.

Assim que a gente soube do incidente, as equipes voltaram o foco para a restauração para todos os clientes afetados. Naquele momento, a gente estimou que o número de sites impactados fosse de aproximadamente 700 (883 sites no total foram afetados, mas a gente subtraiu os sites de propriedade da Atlassian). Dos 700, uma parcela significativa era de contas inativas, gratuitas ou pequenas com um baixo número de usuários ativos. Com base nisso, a gente estimou inicialmente o número aproximado de clientes afetados em cerca de 400.

Agora a gente tem uma visão muito mais precisa e, para total transparência com base na definição oficial de clientes da Atlassian, 775 clientes foram afetados pela interrupção. No entanto, a maioria dos usuários estava representada dentro da estimativa original de 400 clientes. A interrupção durou até 14 dias para um subconjunto desses clientes, com o primeiro conjunto de clientes tendo os sites restaurados em 8 de abril e todos os clientes tendo os sites restaurados em 18 de abril.

Como foi a coordenação

O primeiro ticket de suporte foi criado por um cliente afetado às 07:46 UTC de 5 de abril. O monitoramento interno da Atlassian não detectou nenhum problema porque os sites foram excluídos por meio de um fluxo de trabalho padrão. Às 08:17 UTC, a gente acionou o processo de gerenciamento de incidentes graves, formando uma equipe de gerenciamento de incidentes multifuncional e, em sete minutos, às 08:24 UTC, ele foi escalado para crítico. Às 08:53 UTC, a equipe da Atlassian confirmou que o ticket de suporte ao cliente e a execução do script estavam relacionados. Assim que a gente percebeu a complexidade da restauração, a gente atribuiu o maior nível de gravidade ao incidente às 12:38 UTC.

A equipe de gerenciamento de incidentes era composta por indivíduos de várias equipes da Atlassian, incluindo engenharia, suporte ao cliente, gerenciamento de programas, comunicações e muitas outras. A equipe principal fez reuniões a cada três horas durante o incidente, até que todos os sites fossem restaurados, validados e devolvidos aos clientes.

Para gerenciar o progresso da restauração, a gente criou um novo projeto do Jira, SITE, e um fluxo de trabalho para monitorar restaurações site por site com várias equipes (engenharia, gerenciamento de programas, suporte etc.). Essa abordagem permitiu que todas as equipes identificassem e monitorassem com facilidade problemas relacionados a qualquer restauração de site individual.

A gente também implementou um congelamento de código em toda a engenharia durante o incidente em 8 de abril às 03:30 UTC. Assim a gente pôde manter o foco na restauração do cliente, eliminar o risco de alterações que causassem inconsistências nos dados dos clientes, minimizar o risco de outras interrupções e reduzir a probabilidade de que alterações não relacionadas distraíssem a equipe da recuperação.

Linha do tempo do incidente

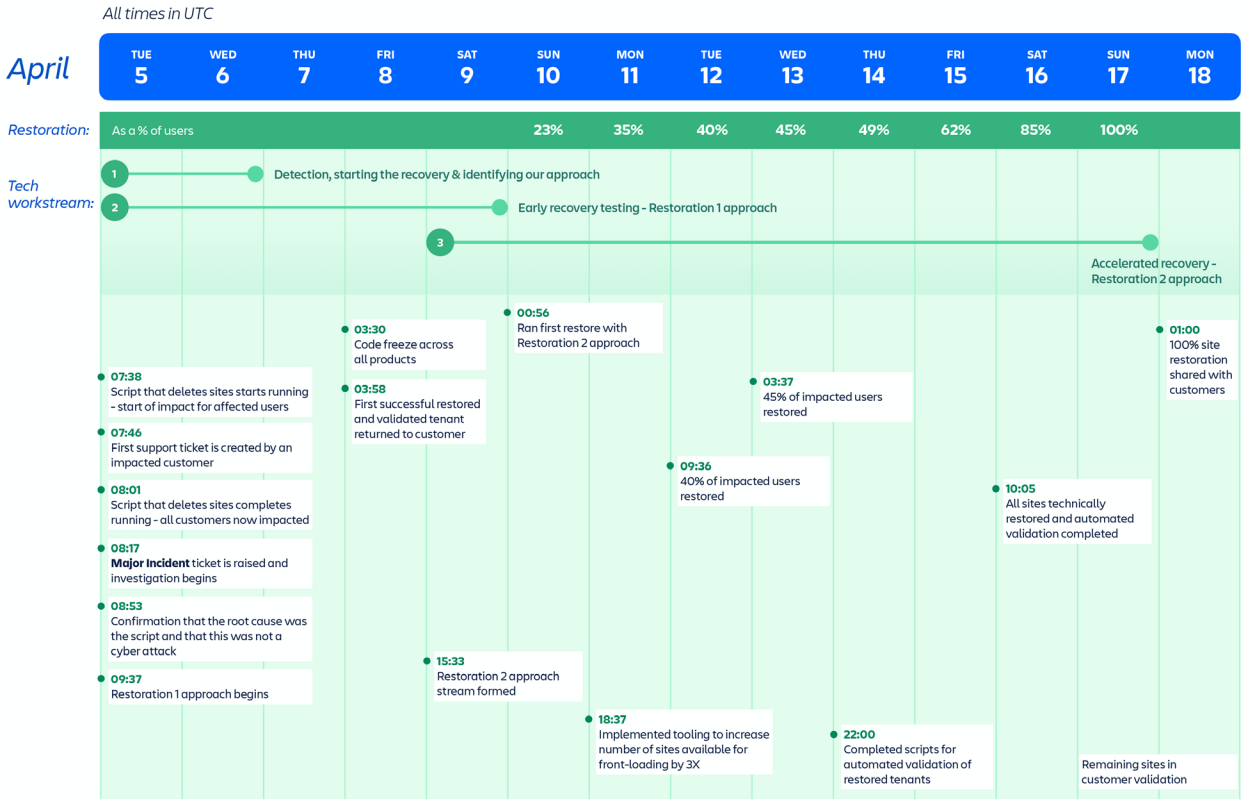


Figura 6: Linha do tempo do incidente e principais marcos da restauração.

Visão geral de alto nível dos fluxos de trabalho de recuperação

A recuperação foi executada como três fluxos de trabalho principais: detecção, recuperação inicial e aceleração. Embora a gente tenha descrito cada fluxo de trabalho em separado abaixo, durante a recuperação havia trabalho sendo feito em paralelo em todos os fluxos de trabalho.

Fluxo de trabalho 1: detecção, início da recuperação e identificação da abordagem

Carimbo de data/hora: dias 1 e 2 (5 e 6 de abril)
Às 08:53 UTC de 5 de abril, a gente identificou que o script do Insight App causou a exclusão de sites. A gente confirmou que não era resultado de um ato mal-intencionado interno nem de ataque cibernético. As equipes relevantes de infraestrutura de plataforma e produto foram chamadas e trazidas para o incidente.

No início do incidente, a gente reconheceu que:

- A restauração de centenas de sites excluídos é um processo complexo de várias etapas (detalhado na seção de arquitetura acima), que exige muitas equipes e vários dias para ser concluído com êxito.
- A gente tinha a capacidade de recuperar um único site, mas não tinha construído recursos e processos para recuperar um grande lote de sites.

Como resultado, a gente precisava paralelizar e automatizar substancialmente o processo de restauração para ajudar os clientes afetados a recuperar o acesso aos produtos Atlassian o mais rápido possível.

O fluxo de trabalho 1 envolveu um grande número de equipes de desenvolvimento trabalhando nas seguintes atividades:

- Identificar e executar etapas de restauração para lotes de sites no pipeline.
- Escrever e melhorar a automação para permitir que a(s) equipe(s) executasse(m) etapas de restauração para um número maior de sites em um lote.

Fluxo de trabalho 2: recuperação inicial e abordagem Restauração 1

Carimbo de data/hora: dias 1 a 4 (5 a 9 de abril)
A gente entendeu o que causou a exclusão dos sites em 5 de abril às 08:53 UTC dentro de uma hora após o script terminar sua execução. A gente também identificou o processo de

restauração que havia sido usado antes para recuperar um pequeno número de sites para produção. No entanto, o processo de recuperação para restaurar sites excluídos em tal escala não estava bem definido.

Para avançar rapidamente, os estágios iniciais do incidente foram divididos entre dois grupos de trabalho:

- O grupo de trabalho manual validou as etapas necessárias e executou manualmente o processo de restauração para um pequeno número de sites.
- O grupo de trabalho de automação pegou o processo de restauração existente e construiu uma automação para executar com segurança as etapas em lotes maiores de sites.

Visão geral da abordagem Restauração 1 (veja a *Figura 7* abaixo):

- Ela exigiu a criação de um site novo para cada site excluído, seguido por cada produto, serviço e armazenamento de dados downstream que precisasse de restauração dos dados.
- O novo site viria com novos identificadores, como o `cloudId`. Esses identificadores são todos considerados imutáveis, o que significa que muitos sistemas incorporam esses identificadores em registros de dados. Como resultado, a gente precisaria atualizar grandes quantidades de dados se esses identificadores mudassem, o que é particularmente problemático para aplicativos de ecossistemas de terceiros.
- Modificar um novo site para replicar o estado do site excluído tinha dependências complexas e muitas vezes imprevistas entre as etapas.

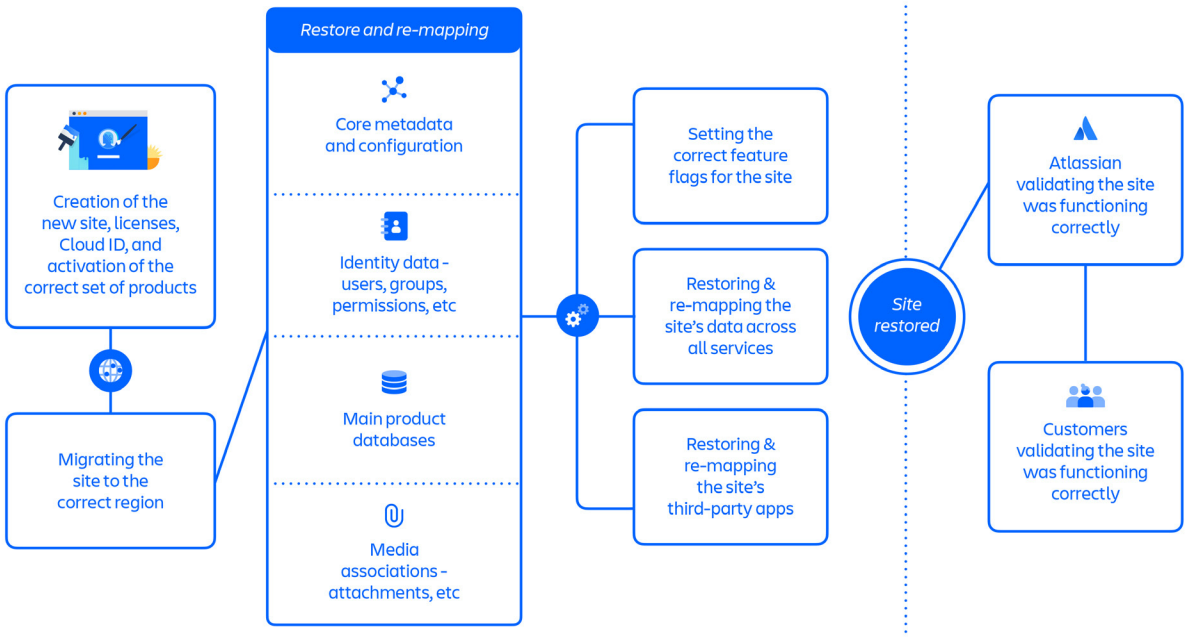


Figura 7: Etapas principais na abordagem Restauração 1.

A abordagem Restauração 1 incluiu aproximadamente 70 etapas individuais que, quando agregadas em um nível alto, seguiram um fluxo amplamente sequencial de:

- Criação de novos sites, licenças, Cloud ID e ativação do conjunto correto de produtos
- Migração dos sites para a região correta
- Restauração e remapeamento dos principais metadados e configuração dos sites
- Restauração e remapeamento dos dados de identidade dos sites – usuários, grupos, permissões, etc.
- Restauração dos principais bancos de dados de produtos dos sites
- Restauração e remapeamento das associações de mídia dos sites – anexos, etc.
- Definição dos sinalizadores de funcionalidade corretos para os sites
- Restauração e remapeamento dos dados dos sites em todos os serviços
- Restauração e remapeamento dos aplicativos de terceiros dos sites
- A validação dos sites pela Atlassian estava funcionando corretamente
- A validação dos sites pelos clientes estava funcionando corretamente

Depois de otimizada, a abordagem Restauração 1 levou aproximadamente 48 horas para restaurar um lote de sites e foi usada para a recuperação de 53% dos usuários afetados em 112 sites entre 5 de abril e 14 de abril.

Fluxo de trabalho 3: recuperação acelerada e abordagem Restauração 2

Carimbo de data/hora: dias 4 a 13 (9 a 17 de abril)
Com a abordagem Restauração 1, a gente levaria três semanas para restaurar os sites de todos os clientes. Portanto, a gente propôs uma nova abordagem em 9 de abril para acelerar a restauração de todos os sites, a abordagem Restauração 2 (veja a *Figura 8* abaixo).

A abordagem Restauração 2 ofereceu paralelismo aprimorado entre as etapas de restauração ao reduzir a complexidade e o número de dependências que estavam presentes com a abordagem Restauração 1.

A Restauração 2 envolveu a recriação (ou a reversão da exclusão) de registros associados aos sites em todos os sistemas respectivos, começando com o registro do Serviço de Catálogo. Um elemento-chave dessa nova abordagem foi *reutilizar todos os identificadores de sites antigos*. Assim foi possível remover mais da metade das etapas do processo anterior que foram usadas para mapear os identificadores antigos para os identificadores novos, incluindo a necessidade de coordenação com todos os fornecedores de aplicativos de terceiros de cada site.

No entanto, a mudança da abordagem Restauração 1 para a Restauração 2 adicionou sobrecarga substancial na resposta a incidentes:

- Muitos dos scripts e processos de automação estabelecidos na abordagem Restauração 1 tiveram que ser modificados para Restauração 2.
- As equipes que realizavam restaurações (incluindo coordenadores de incidentes) precisavam gerenciar lotes paralelos de restaurações em ambas as abordagens, enquanto a gente testava e validava o processo da Restauração 2.
- Usar uma nova abordagem significava que a gente precisava testar e validar o processo da Restauração 2 antes que ele pudesse ser escalado, o que exigia duplicar o trabalho de validação que tinha sido concluído para a Restauração 1.

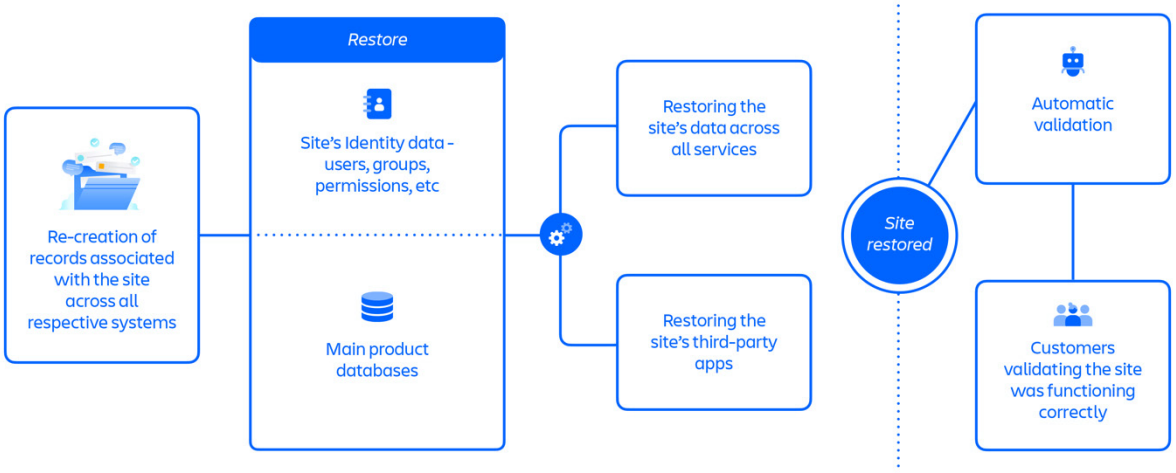


Figura 8: Etapas principais na abordagem Restauração 2.

O gráfico acima representa a abordagem Restauração 2, que incluiu mais de 30 etapas que seguiram um fluxo amplamente paralelizado de:

- Recriação de registros associados aos sites em todos os respectivos sistemas
- Restauração dos dados de identidade dos sites – usuários, grupos, permissões, etc.
- Restauração dos principais bancos de dados de produtos dos sites
- Restauração dos dados dos sites em todos os serviços
- Restauração dos aplicativos de terceiros dos sites
- Validação automática
- A validação dos sites pelos clientes estava funcionando corretamente

Como parte da recuperação acelerada, a gente também tomou medidas para antecipar e automatizar a restauração dos sites, pois a restauração manual não seria bem escalada para grandes lotes. A natureza sequencial do processo de recuperação significava que a restauração dos sites poderia ser mais lenta para restaurações de

bancos de dados grandes e restaurações de permissões e bases de usuários. As otimizações implementadas incluíram:

- A gente desenvolveu as ferramentas e as medidas de proteção necessárias para a *antecipação* e as etapas de longa duração, como restaurações de banco de dados e sincronizações de identidade, para que fossem concluídas antes de outras etapas de restauração.
- As equipes de engenharia criaram automação para suas etapas individuais, o que permitiu que grandes lotes de restaurações fossem executados com segurança.
- Uma automação foi criada para validar o funcionamento correto dos sites após a conclusão de todas as etapas de restauração.

A abordagem acelerada da Restauração 2 levou aproximadamente 12 horas para restaurar um site e foi usada para a recuperação de aproximadamente 47% dos usuários afetados em 771 sites entre 14 e 17 de abril.

Perda mínima de dados após a restauração de sites excluídos

O backup dos bancos de dados da Atlassian é uma combinação de backups completos e backups incrementais que permitem que a gente escolha qualquer "ponto no tempo" específico para recuperar os armazenamentos de dados dentro do período de retenção de backup (30 dias). Para a maioria dos clientes durante este incidente, a gente identificou os principais armazenamentos de dados dos produtos da Atlassian e decidiu usar um ponto de restauração de cinco minutos antes da exclusão dos sites como um ponto de sincronização seguro. Os armazenamentos de dados não primários foram restaurados para o mesmo ponto ou pela reprodução dos eventos registrados. O uso de um ponto de restauração fixo para armazenamentos primários permitiu que a gente tivesse consistência de dados em todos os armazenamentos de dados.

Para 57 clientes que tiveram os sites restaurados no início da resposta ao incidente, a falta de políticas consistentes e a recuperação manual de instantâneos de backup de banco de dados resultaram em alguns bancos de dados do Confluence e do Insight sendo restaurados para um ponto de *mais* de cinco minutos antes da exclusão do site.
A inconsistência foi descoberta durante um processo de auditoria pós-restauração. Desde então, a gente recuperou o restante dos dados, entrou em contato com os clientes afetados por essa situação e está disponibilizando ajuda para aplicar alterações para restaurar ainda mais seus dados.

Em resumo:

- A gente atingiu o Objetivo de Ponto de Recuperação (RPO) de uma hora durante este incidente.

- A perda de dados do incidente foi limitada a cinco minutos antes da exclusão do site.
- Um pequeno número de clientes teve seus bancos de dados do Confluence ou do Insight restaurados para um ponto de mais de cinco minutos antes da exclusão do site. No entanto, a gente pode recuperar os dados e atualmente está trabalhando com os clientes para restaurar esses dados.

Comunicações de incidentes

Quando a gente fala sobre comunicações de incidentes, elas englobam pontos de contato com clientes, parceiros, mídia, analistas do setor, investidores e a comunidade tecnológica mais ampla.

O que aconteceu

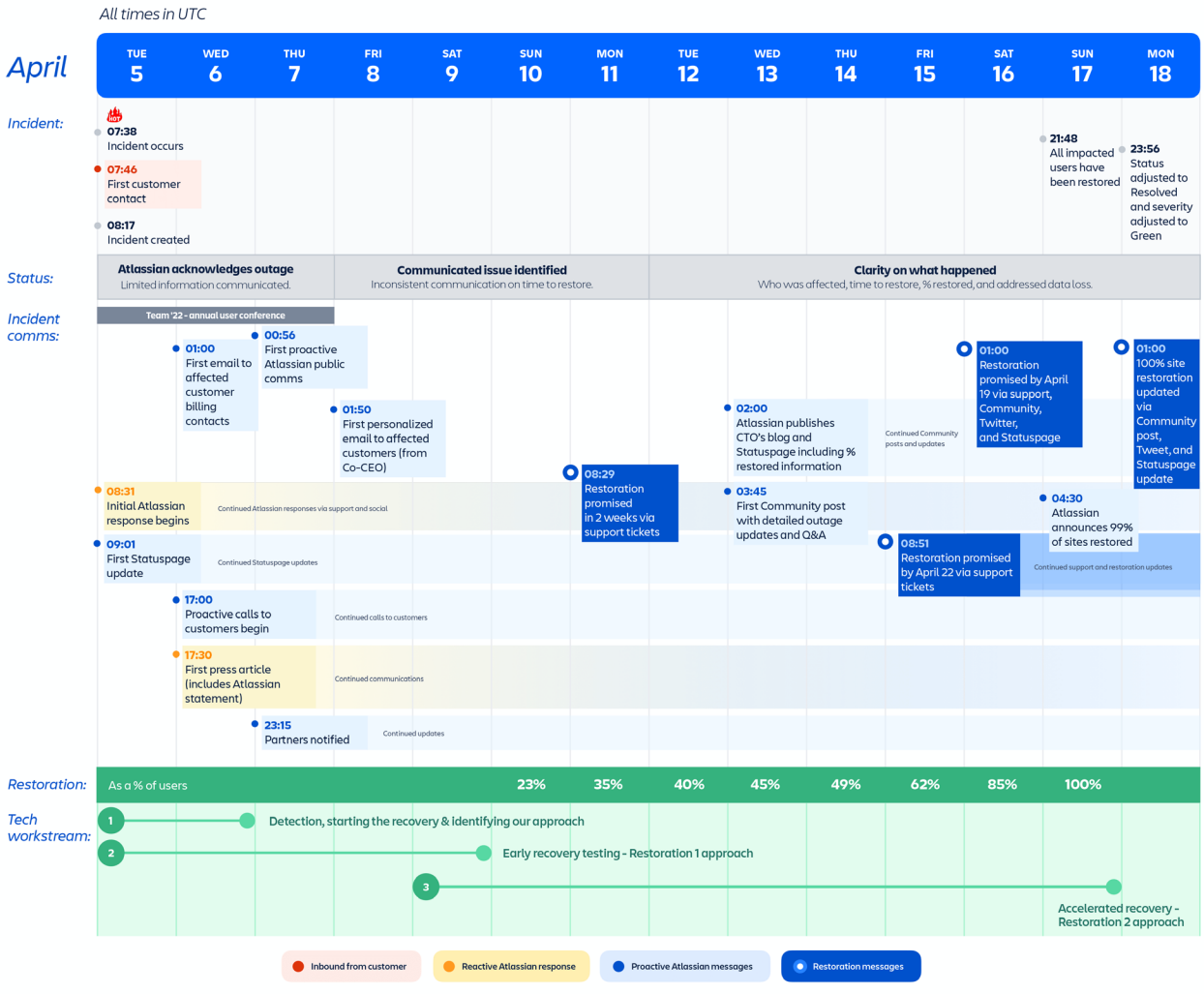


Figura 9: Linha do tempo dos principais marcos das comunicações do incidente.

Carimbo de data/hora: dias 1 a 3 (5 a 7 de abril)

Resposta inicial

O primeiro ticket de suporte foi criado em 5 de abril às 7:46 UTC, e o suporte da Atlassian respondeu reconhecendo o incidente às 8:31 UTC. Às 9:03 UTC, a primeira atualização do Statuspage foi publicada informando aos clientes que a gente estava investigando o incidente. E às 11:13 UTC, a gente confirmou por meio do Statuspage que a gente identificou a causa raiz e estava trabalhando em uma correção. À 1:00 UTC do dia 6 de abril, as comunicações iniciais de tickets de clientes declaravam que a interrupção tinha ocorrido por um script de manutenção e que a gente esperava perda mínima de dados. A Atlassian respondeu às perguntas da mídia com uma declaração em 6 de abril às 17:30 UTC. A Atlassian publicou no Twitter sua primeira mensagem externa ampla reconhecendo o incidente em 7 de abril à 00:56 UTC.

Carimbo de data/hora: dias 4 a 7 (8 a 11 de abril)

Começa um contato mais amplo e personalizado

Em 8 de abril, à 1:50 UTC, a Atlassian enviou um e-mail para os clientes afetados com um pedido de desculpas do cofundador e co-CEO, Scott Farquhar. Nos dias que se seguiram, a gente trabalhou para restaurar as informações de contato excluídas e criar tickets de suporte para todos os sites afetados que ainda não tinham um ticket registrado. A equipe de suporte da Atlassian continuou a enviar atualizações regulares sobre os esforços de restauração por meio dos tickets de suporte associados a cada site afetado.

Carimbo de data/hora: dias 8 a 14 (12 a 18 de abril)

Maior clareza e restauração completa

Em 12 de abril, a Atlassian publicou uma atualização do CTO, Sri Viswanath, disponibilizando mais informações técnicas sobre o que aconteceu, quem foi afetado, se houve perda de dados, o progresso na restauração e dizendo que podia levar até duas semanas para restaurar totalmente todos os sites. O blog foi acompanhado por outra declaração de imprensa atribuída a Sri. A gente também mencionou o blog de Sri na primeira publicação proativa do chefe de engenharia, Stephen Deasy, na Comunidade da Atlassian, que posteriormente se tornou o local dedicado para atualizações adicionais e perguntas e respostas com o público em geral. Uma atualização de 18 de abril desta publicação anunciou a restauração completa de todos os sites de clientes afetados.

Por que não a gente não deu uma resposta pública antes?

1. A gente priorizou a comunicação direta com os clientes afetados via Statuspage, e-mail, tickets de suporte e interações individuais. No entanto, a gente não conseguiu alcançar muitos clientes porque as informações de contato foram perdidas quando os sites foram excluídos. A gente deveria ter implementado comunicações mais amplas muito antes, a fim de informar os clientes e usuários finais afetados sobre a linha do tempo de resposta e resolução de incidentes.

2. Embora a gente soubesse imediatamente o que havia causado o incidente, a complexidade arquitetônica e as circunstâncias únicas deste incidente diminuíram a velocidade da capacidade de definir rapidamente o escopo e estimar com precisão o tempo de resolução. Em vez de esperar até que a gente tivesse uma visão completa, a gente deveria ter comunicado com mais transparência o que sabia e o que não sabia. Informar estimativas gerais de restauração (mesmo que direcionais) e falar com clareza sobre quando a gente esperava ter uma visão mais completa permitiria que os clientes planejassem melhor como lidar com o incidente. Essa situação é ainda mais real para administradores de sistema e responsáveis técnicos, que estão na linha de frente do gerenciamento de partes interessadas e usuários em suas empresas.

Experiência de suporte e contato com o cliente

Como mencionado anteriormente, o mesmo script que excluiu sites de clientes também excluiu os principais identificadores de clientes e informações de contato (por exemplo, URL do Cloud e contatos do administrador do sistema do site) dos ambientes de produção da Atlassian. É um ponto notável porque os sistemas principais da Atlassian (por exemplo, suporte, licenciamento e faturamento) aproveitam a existência de uma URL do Cloud e dos contatos do administrador do sistema do site como identificadores primários para fins de segurança, encaminhamento e priorização. Quando esses identificadores foram perdidos, a gente inicialmente perdeu a capacidade de identificar e interagir com os clientes sistematicamente.

Como o suporte aos clientes foi afetado?

Em primeiro lugar, a maioria dos clientes afetados não conseguiu entrar em contato com a equipe de suporte por meio do formulário de contato on-line normal. Esse formulário foi criado para exigir que o usuário entre com sua ID da Atlassian e informe uma URL do Cloud válida. Sem uma URL válida, o usuário é impedido de enviar um ticket de suporte técnico. No curso normal dos negócios, essa verificação é intencional para a segurança do site e a triagem de tickets. No entanto, esse requisito criou um resultado não intencional para os clientes afetados por essa interrupção: eles foram impedidos de enviar um ticket de suporte ao site de alta prioridade.

Em segundo lugar, a exclusão dos dados do administrador do sistema do site causada pelo incidente criou uma lacuna na capacidade de interação proativa da Atlassian com os clientes afetados. Nos primeiros dias do incidente, a gente enviou comunicações proativas aos responsáveis técnicos e por pagamentos dos clientes afetados registrados na Atlassian. No entanto, a gente logo identificou que muitos responsáveis técnicos e por pagamentos dos clientes afetados estavam desatualizados. Sem as informações de administrador do sistema de cada site, a gente não tinha uma lista completa de contatos ativos e aprovados para com quem interagir.

Como a gente respondeu?

As equipes de suporte da Atlassian tinham três prioridades igualmente importantes para acelerar a restauração dos sites e reparar a quebra nos canais de comunicação nos primeiros dias do incidente.

Primeiro, obter uma lista confiável de contatos de clientes validados. Enquanto as equipes de engenharia trabalhavam para restaurar os sites dos clientes, as equipes voltadas para o cliente mantinham o foco em restaurar informações de contato validadas. A gente usou todos os mecanismos à disposição (sistemas de faturamento, tickets de suporte anteriores, outros backups seguros dos usuários, contato direto com o cliente, etc.) para reconstruir a lista de contatos. O objetivo era ter um ticket de suporte relacionado ao incidente para cada site afetado para agilizar o contato direto e os tempos de resposta.

Segundo, restabelecer fluxos de trabalho, filas e SLAs específicos para este incidente. A exclusão do Cloud ID e a incapacidade de autenticar usuários corretamente também afetaram a capacidade de processar tickets de suporte relacionados ao incidente por meio dos sistemas normais. Os tickets não apareciam corretamente nos painéis e nas filas de escalonamento e prioridade relevantes. A gente logo criou uma equipe multifuncional (suporte, produto e TI) para projetar e adicionar mais lógica, SLAs, estados

de fluxo de trabalho e painéis. Como essas medidas tinham que ser feitas no sistema de produção, levou vários dias para o desenvolvimento, o teste e a implementação totais.

Terceiro, escalar massivamente as validações manuais para acelerar as restaurações dos sites. À medida que a engenharia avançava nas restaurações iniciais, ficou claro que a capacidade das equipes de suporte globais seria necessária para ajudar a acelerar a recuperação dos sites por meio de testes e verificações de validação manuais. Esse processo de validação passaria a ser um caminho crítico para a restauração de sites para os clientes, uma vez que a equipe de engenharia acelerasse as restaurações de dados. A gente teve que criar um fluxo independente de procedimentos operacionais padrão (POPs), fluxos de trabalho, entregas e listas de pessoal para mobilizar mais de 450 engenheiros de suporte para executar verificações de validação, com turnos que ofereciam cobertura 24 horas por dia, 7 dias por semana, para que as restaurações chegassem mais rápido nas mãos dos clientes.

Mesmo com essas prioridades-chave bem estabelecidas ao final da primeira semana, a capacidade de disponibilizar atualizações *significativas* estava limitada pela falta de clareza sobre as linhas do tempo de resolução de incidentes pela complexidade dos processos de restauração. A gente deveria ter reconhecido a incerteza em informar uma data de restauração dos sites mais cedo e se disponibilizado mais cedo para discussões pessoais, para que os clientes pudessem fazer planos de acordo com o contexto.

Como a gente vai melhorar?

A gente bloqueou imediatamente as exclusões de sites em massa até que as alterações adequadas possam ser feitas.

Neste momento em que a gente supera este incidente e reavalia os processos internos, a gente quer reconhecer que pessoas não causam incidentes. Na verdade, os sistemas permitem que erros sejam cometidos. Esta seção resume os fatores que contribuíram para este incidente. A gente também discute os planos para acelerar a forma como vai corrigir essas fraquezas e esses problemas.

Aprendizado 1: as “exclusões temporárias” devem ser universais em todos os sistemas

No geral, uma exclusão desse tipo deve ser proibida ou ter várias camadas de proteção para evitar erros. A principal melhoria que a gente está fazendo é impedir globalmente a exclusão de dados e metadados dos clientes que não passaram por um processo de exclusão temporária.

a) A exclusão de dados deve ocorrer apenas como exclusão temporária
A exclusão de um site inteiro deve ser proibida, e a exclusão temporária deve exigir proteções em vários níveis para evitar erros. A gente vai implementar uma política de "exclusão temporária", impedindo que scripts ou sistemas externos excluam dados dos clientes em um ambiente de produção. A política de "exclusão temporária" vai permitir retenção de dados suficiente para que a recuperação de dados possa ser executada com rapidez e segurança. Os dados só vão ser excluídos do ambiente de produção depois que um período de retenção expirar.

Ações:

- **Implementar uma "exclusão temporária" nos fluxos de trabalho de provisionamento e em todos os armazenamentos de dados relevantes:** além disso, a equipe da plataforma de locatários vai verificar se as exclusões de dados só podem ocorrer após as desativações, bem como outras salvaguardas nesse espaço. A longo prazo, a plataforma de locatários vai assumir um papel de liderança para desenvolver ainda mais o gerenciamento correto do estado dos dados do locatário.

b) A exclusão temporária deve ter um processo de revisão padronizado e verificado
As ações de exclusão temporária são operações de alto risco. Como tal, a gente deve ter processos de revisão padronizados ou automatizados que incluam rollbacks definidos e procedimentos de teste para lidar com essas operações.

Ações:

- **Implementação em etapas obrigatória para qualquer ação de exclusão temporária:** todas as novas operações que exigem exclusão vão ser primeiro testadas em nossos próprios sites para validar a abordagem e verificar a automação. Depois de concluída essa validação, a gente vai fazer a migração progressiva dos clientes pelo mesmo processo e continuar fazendo testes para ver se há irregularidades antes de aplicar a automação a toda a base de usuários selecionada.

- **As ações de exclusão temporária devem ter um plano de rollback testado:** qualquer atividade de exclusão temporária de dados deve testar a restauração dos dados excluídos antes de ser executada na produção e ter um plano de rollback testado.

Aprendizado 2: como parte do programa de recuperação de desastres, automatizar a restauração para eventos de exclusão de vários sites e vários produtos para um conjunto maior de clientes

O Gerenciamento de Dados da Atlassian descreve o processo de gerenciamento de dados da Atlassian em detalhes. Para oferecer alta disponibilidade, a gente provisiona e mantém uma réplica síncrona em espera em várias Zonas de Disponibilidade (AZs) da AWS. O failover da AZ é automatizado e, via de regra, leva de 60 a 120 segundos, e a gente lida regularmente com interrupções de data center e outras interrupções comuns sem impacto para os clientes.

Também mantemos backups imutáveis que são projetados para serem resilientes contra eventos de corrupção de dados, o que permite a recuperação para um ponto anterior no tempo. Os backups são mantidos por 30 dias, e a Atlassian testa e audita sempre os backups de armazenamento para restauração. E, se necessário, a gente pode restaurar todos os clientes de uma só vez em um novo ambiente.

Usando esses backups, a gente faz rollbacks regulares de clientes individuais ou um pequeno conjunto de clientes que excluem seus próprios dados por acidente. No entanto, a exclusão no nível do site não tinha runbooks que pudessem ser rapidamente automatizados para a escala desse evento, que exigia que ferramentas e automação em todos os produtos e serviços funcionassem de forma coordenada.

O que ainda não automatizamos (ainda) é restaurar um grande subconjunto de clientes no ambiente existente (e em uso atual) sem afetar nenhum dos outros clientes.

No ambiente de nuvem, cada armazenamento de dados contém dados de vários clientes. Como os dados excluídos neste incidente eram apenas uma parte dos armazenamentos de dados que continuam sendo usados por outros clientes, a gente precisa fazer a extração e a restauração manuais de partes individuais dos backups. Cada recuperação de site do cliente é um processo demorado e complexo, que exige validação interna e verificação final do cliente quando o site é restaurado.

Ações:

- **Acelerar restaurações de vários produtos e vários sites para um conjunto maior de clientes:** o programa de DR atende aos padrões atuais de RPO de uma hora. A gente vai aproveitar a automação e os aprendizados deste incidente para acelerar o programa de DR para atender ao RTO, conforme definido na política para essa escala de incidente.

Automatizar e adicionar a verificação desse caso ao teste de DR: a gente vai executar regularmente exercícios de DR que envolvam a restauração de todos os produtos para um grande conjunto de sites. Esses testes de DR vão verificar se os runbooks estão atualizados conforme a arquitetura evolui e se novos casos extremos forem encontrados. A gente vai sempre melhorar a abordagem de restauração, automatizar mais o processo de restauração e reduzir o tempo de recuperação.

Aprendizado 3: melhorar o processo de gerenciamento de incidentes para eventos de grande escala

O programa de gerenciamento de incidentes da Atlassian é adequado para gerenciar os incidentes graves e leves que ocorreram ao longo dos anos. A gente frequentemente simula a resposta a incidentes para incidentes de menor escala e duração mais curta, que costumam envolver menos pessoas e equipes.

No entanto, em seu auge, este incidente teve centenas de engenheiros e funcionários de suporte ao cliente trabalhando simultaneamente para restaurar os sites dos clientes. O programa de gerenciamento de incidentes e as equipes da Atlassian não foram projetados para lidar com a profundidade, a amplitude e a duração desse tipo de incidente (consulte a *Figura 10* abaixo).

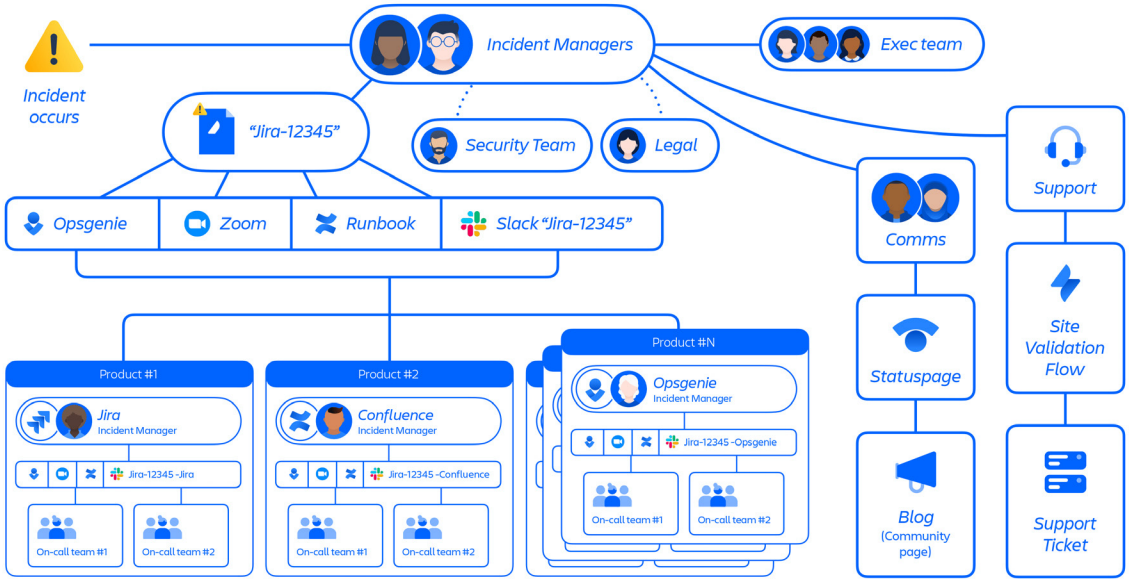


Figura 10: Visão geral do processo de gerenciamento de incidentes em grande escala.

O processo de gerenciamento de incidentes em grande escala vai ser mais bem definido e praticado com frequência

A gente tem esquemas táticos para incidentes no nível do produto, mas não para eventos dessa escala, com centenas de pessoas trabalhando ao mesmo tempo em toda a empresa. Nas ferramentas de gerenciamento de incidentes, a gente tem automação que cria fluxos de comunicação como Slack, Zoom e documentos do Confluence, mas falta a criação de subfluxos que são necessários para incidentes em grande escala para isolar fluxos de restauração.

Ações:

- **Definir um esquema tático e ferramentas para incidentes de grande escala e realizar exercícios simulados:** definir e documentar os tipos de incidentes que podem ser considerados de grande escala e exigem esse nível de resposta. Descrever as principais etapas de coordenação e criar ferramentas para ajudar os gerenciadores de incidentes e outras funções de negócios a agilizar a resposta e iniciar a recuperação. Os gerenciadores de incidentes e as equipes vão realizar regularmente simulações, treinamentos e o refinamento de ferramentas e documentos para a melhoria contínua.

Aprendizado 4: melhorar os processos de comunicação

a) A gente excluiu identificadores críticos de clientes, impactando as comunicações e ações para as pessoas afetadas

O mesmo script que excluiu sites de clientes também excluiu os principais identificadores dos clientes (por exemplo, URL do site e contatos do administrador do sistema do site) dos ambientes de produção da Atlassian. Como resultado, (1) os clientes foram impedidos de abrir tickets de suporte técnico por meio do canal de suporte normal; (2) a gente dias para obter uma lista confiável dos principais contatos dos clientes (como administradores de sistema do site) afetados pela interrupção para engajamento proativo; e (3) fluxos de trabalho de suporte, SLAs, painéis e os processos de escalonamento não apresentaram o funcionamento adequado inicialmente por conta da natureza única do incidente.

Durante a interrupção, os escalonamentos de clientes também ocorreram por meio de vários canais (e-mail, telefonemas, tickets de CEO, LinkedIn e outros canais sociais, além de tickets de suporte). O uso de ferramentas e processos diferentes entre as equipes voltadas para o cliente diminuíram a velocidade de resposta e dificultaram o monitoramento e a geração de relatórios holísticos desses escalonamentos.

b) A gente não tinha um esquema tático de comunicação de incidentes completo o suficiente para lidar com esse nível de complexidade

A gente não tinha um esquema tático de comunicação de incidentes que delineasse princípios, bem como funções e responsabilidades, para mobilizar uma equipe de comunicação de incidentes unificada e multifuncional rápido o suficiente. A gente não reconheceu o incidente de forma rápida e consistente em meio de vários canais, especialmente nas mídias sociais. Divulgar comunicações públicas mais amplas em torno da interrupção, em conjunto com a repetição da mensagem crítica de que não houve perda de dados e de que a situação não era resultado de um ataque cibernético, teria sido a abordagem correta.

Ações:

- **Melhorar o backup dos principais contatos:** fazer backup das informações de contato autorizadas da conta fora da instância do produto.
- **Ajustar as ferramentas de suporte:** criar mecanismos para clientes sem URL de site válida ou ID da Atlassian para fazer contato direto com a equipe de suporte técnico da Atlassian.
- **Sistema e processos de escalonamento de clientes:** investir em um sistema de escalonamento unificado baseado na conta, além de fluxos de trabalho que permitam que vários objetos de trabalho (tickets, tarefas, etc.) sejam armazenados sob um único objeto da conta do cliente, para melhorar a coordenação e a visibilidade em todas as equipes voltadas para o cliente.
- **Agilizar a cobertura do gerenciamento de escalonamento 24 horas por dia, 7 dias por semana:** executar planos de expansão de presença global para a função de gerenciamento de escalonamento para permitir uma cobertura consistente 24 horas por dia, 7 dias por semana, com funcionários designados com base em cada região geográfica principal, em conjunto com funções de suporte para oferecer assistência com especialistas no assunto e lideranças de vendas e produtos necessários.
- **Atualizar o esquema tático de comunicação de incidentes com novos aprendizados e revisá-lo regularmente:** revisar o esquema tático para definir funções e linhas de comunicação claras internamente. Usar a estrutura DACI para incidentes e ter substitutos 24 horas por dia, 7 dias por semana para cada função em caso de doença, férias ou outros eventos imprevistos. Realizar uma auditoria trimestral para verificar a prontidão em todos os momentos.

Ações (cont.)

Seguir o template de comunicação de incidentes em todas as comunicações: falar sobre o que aconteceu, quem foi afetado, linha do tempo para restauração, porcentagens de restauração dos sites, perda de dados esperada, com os níveis de confiança associados, além de orientações claras sobre como entrar em contato com o suporte.

Considerações finais

Enquanto a interrupção é resolvida e os clientes têm a restauração total, trabalho da Atlassian continua. Neste estágio, a gente está implementando as mudanças descritas acima para melhorar os processos, aumentar a resiliência e evitar que uma situação como essa aconteça de novo.

A Atlassian é uma empresa em constante aprendizado, e as equipes certamente aprenderam muitas lições difíceis com essa experiência. A gente está colocando essas lições em prática para fazer mudanças duradouras nos negócios. Em última análise, a gente vai ficar mais forte e oferecer um serviço melhor por causa dessa experiência.

A gente espera que o aprendizado deste incidente seja útil para outras equipes que fazem um trabalho diligente para oferecer serviços confiáveis a seus clientes.

Por fim, quero agradecer àqueles que estão lendo esta mensagem e aprendendo com a gente e àqueles que fazem parte da comunidade e da equipe da Atlassian.

-Sri Viswanath, CTO

©2022 Atlassian. Todos os direitos reservados. ENTMIG-245 29/04/22

ATLASSIAN