ATLASSIAN

事后回顾

2022 年 4 月中断

仅为方便起见，我们提供了此翻译版本。如果中英文之间存在任何歧义或冲突，以原始英文版本为准。

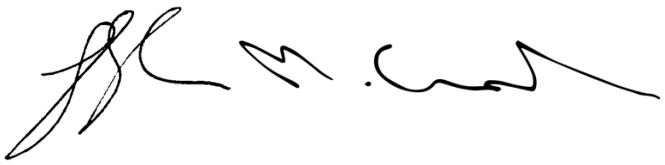
联合创始人和联合首席执行官致信

我们承认本月初发生了客户服务中断事件。我们了解，我们的产品对于您的业务至关重要，因此，我们不会推卸责任。所有责任都由我们承担，完全由我们负责。对于受影响的客户，我们正在努力重新赢得您的信任。

Atlassian 的核心价值观之一就是“开放的公司，绝无虚言”。我们通过公开讨论事件并视其为学习机会来践行这一价值观。我们向客户、Atlassian 社区和规模更大的技术社区发布了这篇事后回顾。我们以自己的事件管理流程为傲，该流程强调建立无责备文化，专注于寻找改进技术系统和流程的方法，这对于提供大规模、值得信赖的服务来说至关重要。虽然我们会竭尽全力避免各类事件的发生，但我们也承认：事件也是推动我们改进的一个强有力方式。

请放心，借助 Atlassian 云平台，我们可满足超过 20 万个不同规模、不同行业云客户的不同需求。在此次事件之前，我们的云一直保持着 99.9% 的正常运行时间，超过了 SLA 的正常运行时间要求。我们还针对我们的平台和一些集中式平台功能进行了长期投资，搭载可扩展的基础架构，同时定期改善安全性。

对于我们的客户和合作伙伴，感谢您一直以来的信任与合作。希望通过本文所述详情和行动，我们能够表明 Atlassian 将继续提供世界一流的云平台和强大的产品组合，
以满足各个团队的需求。



-Scott 和 Mike

执行摘要

自世界标准时间 2022 年 4 月 5 日星期二 7:38 起，775 个 Atlassian 客户失去了 Atlassian 产品的访问权限。其中部分客户的中断时间长达 14 天，第一批客户于 4 月 8 日恢复，所有客户站点都在 4 月 18 日前恢复。

此次事故并非由网络攻击造成，客户数据也未遭到未经授权的访问。Atlassian 设有全面的数据管理计划，发布了 SLA，并且保持着超出 SLA 要求的记录。

尽管这次属于重大事件，但所有客户丢失的数据都不超过五分钟。此外，超过 99.6% 的客户和用户还在继续使用我们的云产品，恢复期间也没有中断。

> 在本文中，我们将在此次事件中遭遇站点删除的客户称为“受影响”客户。本 PIR 披露了确切的事件详情，概述了我们的恢复步骤，并描述了我们如何防止类似事件再次发生。我们在本节提供了关于该事件的高级概述，更多详情见本文剩余部分。

情况说明

2021 年，我们完成了对适用于 Jira Service Management 和 Jira Software 的独立 Atlassian 应用（名为“Insight – Asset Management”）的收购和集成。随后，该独立应用成为 Jira Service Management 中的原生功能，不再适用于 Jira Software。因此，我们需要删除客户站点上安装的旧版独立应用。我们的工程团队使用了现有脚本和程序来删除此独立应用的实例，但存在两个问题：

- **沟通缺失**。请求删除的团队与运行删除的团队之间缺乏沟通。团队没有提供标记为删除的*目标应用*的 ID，而是提供了待删除应用所在*整个云站点*的 ID。

- **系统警示不足**。用于执行删除的 API 同时接受了站点和应用 ID，并假定输入正确，这意味着，如果传递的是站点 ID，将会删除站点；如果传递的是应用 ID，则将删除应用。没有提供警示信号来确定请求删除的类型（是站点还是应用）。

所执行的脚本遵循了我们的标准同行审查程序，该程序主要关注具体调用了哪个端点以及如何调用，但没有反复核对所提供的云站点 ID，没有验证其对应的是 Insight App 还是整个站点，而问题就在于该脚本包含的是客户整个站点的 ID。最终导致在世界标准时间 2022 年 4 月 5 日星期二 07:38 至 08:01 直接删除了 883 个站点（775 个客户）。*参见“情况说明”*

我们是怎样应对的？

在世界标准时间 4 月 5 日 08:17 确认事件后，我们启动了重大事件管理流程，并组建了跨职能事件管理团队。在事件发生期间，这支全球事件响应团队全天候工作，直至所有站点都完成恢复、验证并返还给客户。此外，事件管理负责人每三个小时开一次会，协调相关工作流。

最初，我们发现同时恢复数百个产品各异的客户面临着诸多挑战。

事件发生之初，我们很清楚哪些站点受到了影响，我们的首要任务是与每个受影响站点的获准所有者联系，告知其中断情况。

然而，有些客户的联系信息被删除了。这意味着这些客户无法像往常一样提交支持请求单，同时也意味着我们无法立即联系到关键客户联系人。*更多详情，请参见“恢复工作流*的高级概述”。

我们采取了哪些措施来防止未来再次发生这种情况？

我们立即采取了很多措施，并致力于做出改变，以避免未来再次发生这种情况。下面是我们已经或将要做出重大改变的四个具体领域：

1. **在所有系统中建立通用“软删除”。**总的来说，事件中的这类删除应该被禁止，或受到多层保护，以避免出错，包括针对“软删除”进行分阶段回滚和测试回滚计划。我们将在全球范围内防止跳过软删除程序删除客户数据和元数据。

2. **投资制定我们的灾难恢复 (DR) 计划，为更多客户实现多站点、多产品删除事件自动恢复程序。**我们将利用自动化以及从此次事件中吸取的教训，加快灾难恢复计划，以达到我们在政策中就这一规模事件而设定的恢复时间目标 (RTO)。我们将定期进行灾难恢复演习，包括恢复大批站点的所有产品。

3. **修改大规模事件的事件管理流程。**我们将改进大规模事件的标准操作程序，并通过模拟大规模事件来练习这种程序，我们将会更新我们的培训和工具，以应对大量团队并行工作的情况。

4. **创建大规模事件通信行动手册。**我们将通过多个渠道及早发现事件，并在数小时内发布事件公告。为了更好地接触受影响的客户，我们将改进关键联系人备份并改造支持工具，让客户没有有效的 URL 或 Atlassian ID 也能直接联系我们的技术支持团队。

我们完整的行动项目列表详见下面的完整事后回顾。*参见“我们如何改进”*

目录

Atlassian 的云架构概述

要了解本文档中所述事件的起因，建议先了解一下 Atlassian 产品、服务和基础架构的部署架构。

Atlassian 的云托管架构

Atlassian 使用 Amazon Web Services (AWS) 作为云服务提供商，并在全球多个地区使用其高可用性数据中心设施。每个 AWS 区域都设有一个独立的地理位置，并且分为多个相互隔离，地理上相互分开的数据中心组，也称可用区域 (AZ)。

我们利用 AWS 的计算、存储、网络和数据服务来构建我们的产品和平台组件，因而能够利用 AWS 提供的冗余功能，如可用区域和区域。

分布式服务架构

借助此 AWS 架构，我们对我们解决方案中使用的很多平台和产品服务都进行了托管。其中包括跨多个 Atlassian 产品共享和使用的平台功能，如媒体、身份、商务、我们的编辑器等体验，以及一些产品专用功能，如 Jira 事务服务和 Confluence 分析。

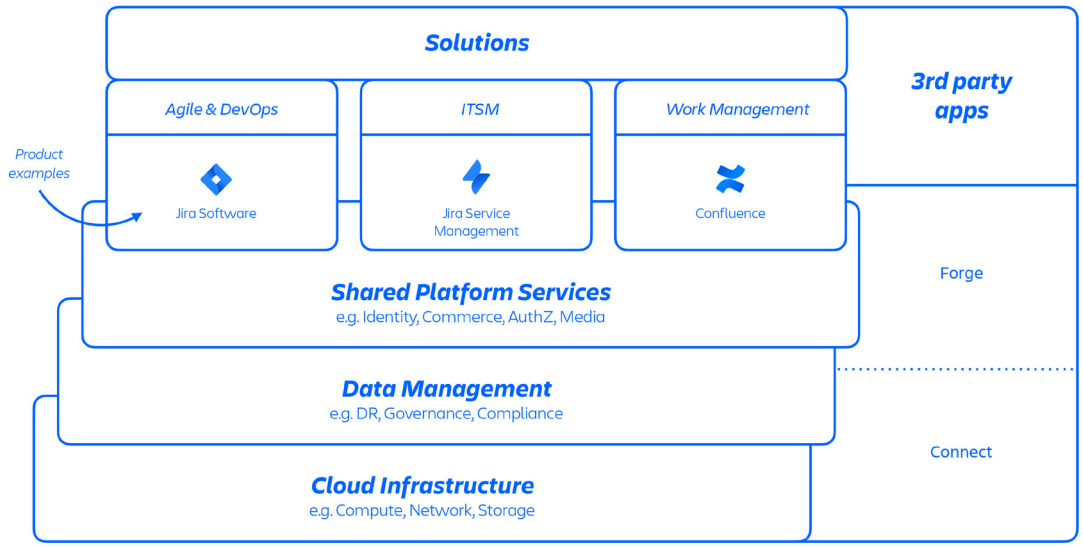


图1：Atlassian 平台架构。

Atlassian 开发人员通过内部开发的 Micros 平台即服务 (PaaS) 来调配这些服务，该平台即服务可自动协调共享服务、基础架构、数据存储及其管理功能的部署，包括安全性和合规性控制要求（见*上图1*）。通常，Atlassian 产品由多个“容器化”服务组成，这些服务通过 Micros 部署在 AWS 上。Atlassian 产品使用的核心平台功能（见 *下图2*）包括请求路由、二进制对象存储、身份验证/授权、事务性用户生成内容 (UGC) 和实体关系存储、数据湖、通用日志、请求跟踪、可观察性和分析服务。这些微服务通过经批准的平台级标准化技术堆栈构建：

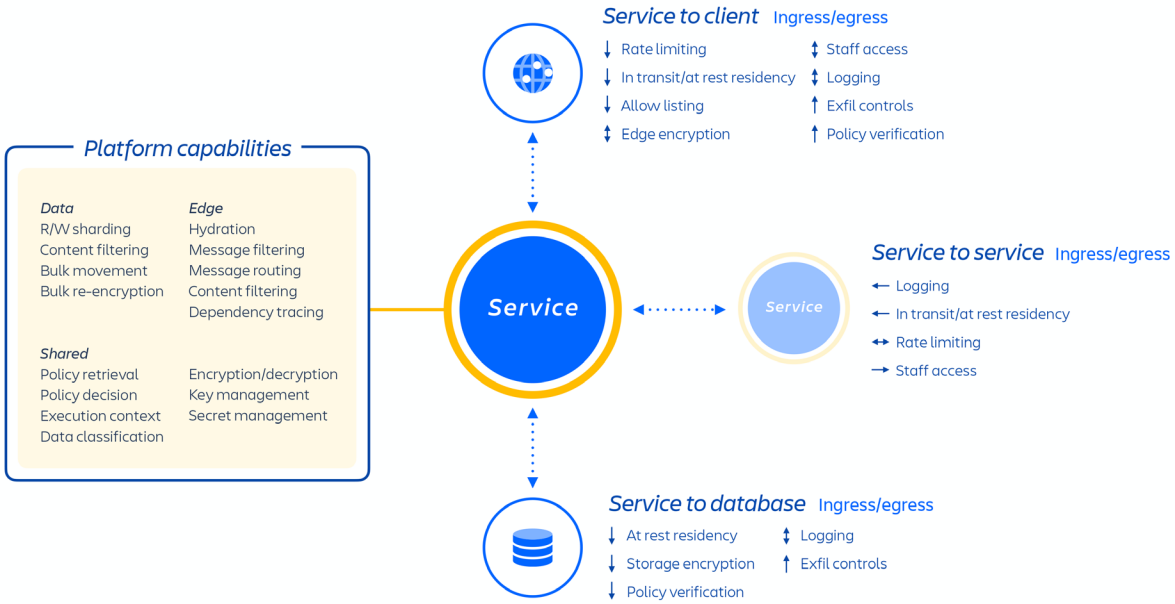


图2：Atlassian 微服务概述。

多租户架构

基于我们的云基础架构，我们构建并运行了一个多租户微服务架构，并带有一个支持我们产品的共享平台。在多租户架构中，单个服务可服务多个客户，包括运行我们的云产品所需的数据库和计算实例。每个分片（本质上是一个容器 - 见 *下图3*）均包含多个租户的数据，但每个租户的数据都是相互隔离的，对其他租户不可见。

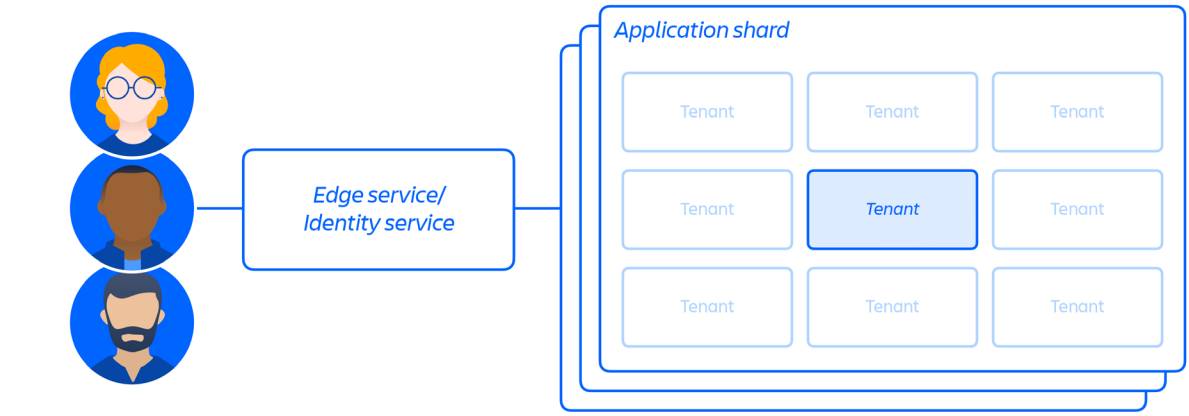


图 3：我们如何在多租户架构中存储数据。

租户调配和生命周期

调配新客户后，很多事件都会触发分布式服务的编排和数据存储的调配。这些事件通常可映射到生命周期中的七个步骤之一：

1. 商务系统会立即更新相应客户的最新元数据和访问控制信息，然后调配编排系统会通过各种租户和产品事件让“已调配资源的状态”与许可状态保持一致。

租户事件

这些事件会影响到整个租户，具体影响可能是：

- 创建：创建租户并用于全新站点
- 销毁：删除整个租户

产品事件

- 激活：激活许可产品或第三方应用后
- 停用：停用特定产品或应用后
- 暂停：在暂停给定的现有产品后，从而收回其对所拥有站点的访问权限
- 取消暂停：在取消暂停给定的现有产品后，从而授予其对所拥有站点的访问权限

许可更新：包含关于给定产品许可席位数及其状态（活动/非活动）的信息

2. 创建客户地点并为客户激活正确的产品集。站点的概念是指许可给特定客户的多种产品的容器。（例如面向 <site-name>.atlassian.net 的 Confluence 和 Jira Software）。这点（见 *下图 4*）对理解本报告上下文至关重要，因为本次事件中删除的就是站点容器，并且本文档自始至终都在讨论站点的概念。

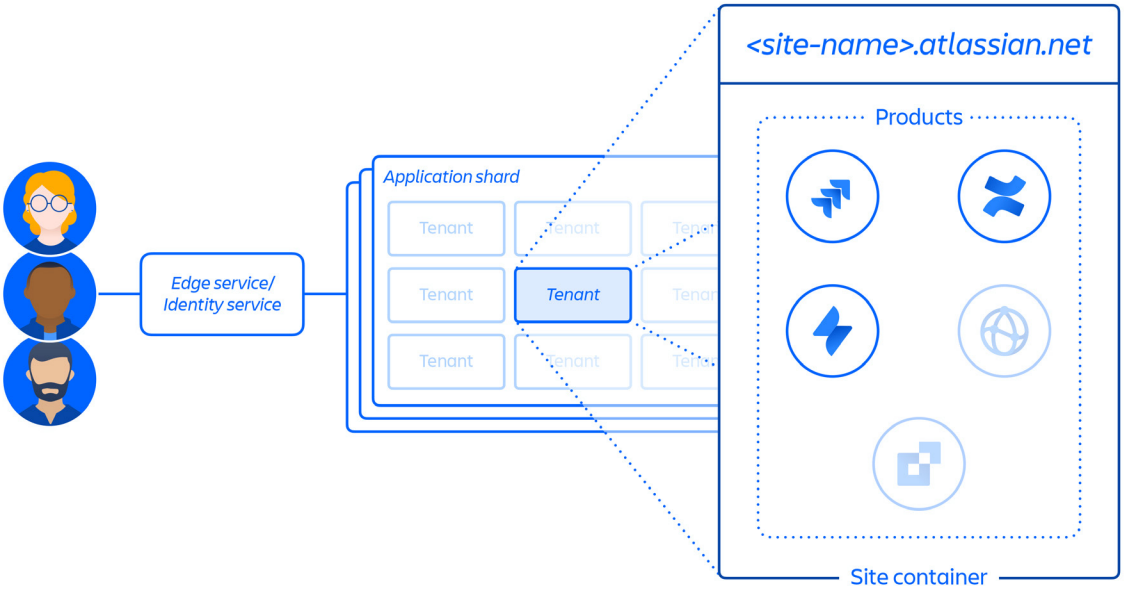


图 4：站点容器概述。

3. 在指定区域的客户站点内调配产品。

   调配产品后，其大部分内容都将托管到靠近用户访问地的位置。为了优化产品性能，我们不会限制在全球托管的数据的移动，我们可能会根据需要在区域之间移动数据。

   对于我们的一些产品，我们还会提供数据驻留功能。通过数据驻留功能，客户可以选择将产品数据分布到全球，或者保存在我们指定的地理区域。

4. 创建并存储客户站点和产品核心元数据和配置。

5. 创建和存储站点和产品标识数据，如用户、组、权限等。

6 在站点内调配产品数据库，例如 Jira 系列产品，Confluence、Compass、Atlas。

7 调配产品许可应用。

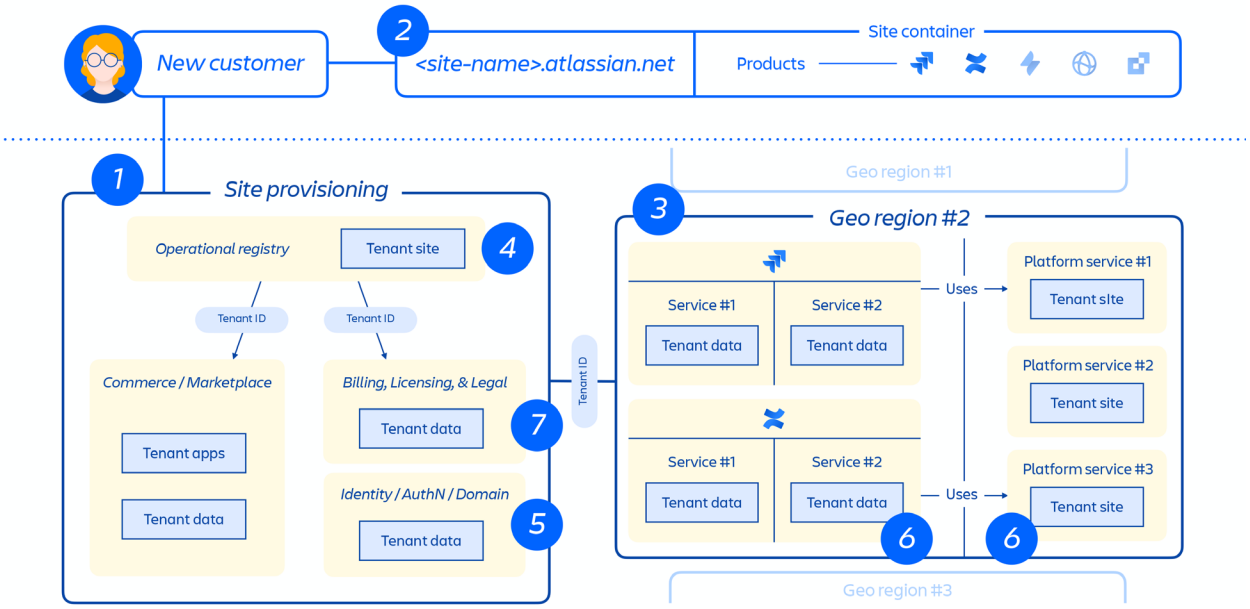


图 5：概述如何在我们的分布式架构中调配客户站点。

上图 5 展示了客户站点在我们的分布式架构中（而不仅仅是在单个数据库或存储中）的部署方式。其中包括存储元数据、配置数据、产品数据、平台数据和其他相关站点信息的多个物理和逻辑位置。

灾难恢复计划

我们的灾难恢复 (DR) 计划涵盖我们为灵活抵御基础架构故障和从备份恢复服务存储的所有付出。要了解灾难恢复计划，先来看看这两个重要概念：

- **恢复时间目标 (RTO)**：如何在灾难期间快速恢复数据并返还给客户？
- **恢复点目标 (RPO)**：从备份中恢复的数据是否为最新数据？自上次备份以来，将会丢失多少数据？

在这起事件中，我们没能达到恢复时间目标 (RTO)，但是达到了恢复点目标 (RPO)。

弹性

我们会为基础架构级别故障做好准备；例如，丢失整个数据库、服务或 AWS 可用区域。此项准备工作包括跨多个可用区域复制数据和服务，以及定期进行故障转移测试。

服务存储可恢复性

我们还准备为因勒索软件、不良行为者、软件缺陷和操作错误等风险造成服务存储数据损坏的情况提供恢复服务。此项准备包括设置不可变备份和服务存储备份恢复测试。我们能够获取任意单个数据存储，并将其恢复到之前的时间点。

多站点、多产品自动恢复能力

此次事件发生时，我们还不具备选择大量客户站点，并将其所有互联产品从备份恢复到之前时间点的能力。

我们的能力都集中在了基础架构、数据损坏、单个服务事件或单站点删除上。过去，我们必须要处理并测试这类故障。站点级删除没有可快速自动化的运行手册，而此次事件的规模需要以协调的方式对所有产品和服务进行工具化和自动化。

下面的章节将更深入地介绍这种复杂性，以及我们在 Atlassian 中为发展和优化大规模维护这一架构的能力而采取的行动。

情况说明、时间线和恢复

情况说明

2021 年，我们完成了对适用于 Jira Service Management 和 Jira Software 的独立 Atlassian 应用（名为“Insight – Asset Management”）的集成。随后，该独立应用成为 Jira Service Management 中的原生功能，不再适用于 Jira Software。因此，我们需要删除客户站点上安装的旧版独立应用。我们的工程团队使用了现有脚本和程序来删除此独立应用的实例。

但由此引发了两个重大问题：

- **沟通缺失。**请求删除的团队与运行删除的团队之间缺乏沟通。团队没有提供标记为删除的目标应用的 ID，而是提供了待删除应用所在整个云站点的 ID。
- **系统警示不足。**用于执行删除的 API 同时接受了站点和应用 ID，并假定输入正确，这意味着，如果传递的是站点 ID，将会删除站点；如果传递的是应用 ID，则将删除应用。没有提供警示信号来确定请求删除的类型（是站点还是应用）。

所执行的脚本遵循了我们的标准同行审查程序，该程序主要关注具体调用了哪个端点以及如何调用，但没有反复核对所提供的云站点 ID，没有验证其对应的是应用还是整个站点。我们根据标准变更管理流程，在 Staging 中对脚本进行了测试，却发现其不会检测到输入的 ID 是不正确的，因为 ID 在 Staging 环境中不存在。

在生产环境中运行时，该脚本最初针对 30 个站点运行。首次生产运行非常成功，删除了 30 个站点的 Insight 应用，且没有产生副作用。但是，这 30 个站点的 ID 是在沟通不畅事件之前获得的，其中包含正确的 Insight 应用 ID。

后续生产运行的脚本包含站点 ID，替换了 Insight 应用 ID，并针对一组 883 个站点执行。该脚本于世界标准时间 4 月 5 日 07:38 开始运行，并于世界标准时间 08:01 完成。脚本根据输入列表按顺序删除了站点，因此在世界标准时间 07:38 脚本开始运行后不久，就删除了第一个客户站点。由此造成 883 站点很快被删除，并且没有向我们的工程团队发布警示信号。

受影响的客户无法使用以下 Atlassian 产品：Jira 系列产品、Confluence、Atlassian Access、Opsgenie 和 Statuspage。

得知事件发生后，我们的团队第一时间开始帮助受影响客户恢复服务。当时，我们估计受影响站点数约为 700 个（实际共计有 883 个站点受到影响，但我们减去了 Atlassian 的自有站点）。在 700 个帐户中，有很大一部分是活跃用户数量较少的非活跃帐户、免费帐户或小型帐户。基于此，我们最初估计受影响的客户大约为 400 人。

现在，我们对具体情况有了更准确的了解，为了根据 Atlassian 的官方客户准则做到公开透明，共有 775 个客户受到了中断影响，大部分用户都属于最初估计的 400 个客户。其中部分客户的中断时间长达 14 天，4 月 8 日为首批客户恢复了服务，4 月 18 日为所有客户恢复了服务。

我们如何协调

第一张支持请求单由一个受影响客户于世界标准时间 4 月 5 日 07:46 创建。我们的内部监控没有检测到问题，因为这些站点都是通过标准工作流删除的。在世界标准时间 08:17，我们启动了重大事件管理流程，组建了一支跨职能事件管理团队，7 分钟后，即世界标准时间 08:24，事件严重程度上升到危急。在世界标准时间 08:53，我们的团队确认客户支持请求单和脚本运行相关。我们意识到恢复复杂性后，随即于世界标准时间 12:38 将事件严重程度标注为最严重。

事件管理团队由来自 Atlassian 多个团队的人员组成，包括工程、客户支持、项目管理、通信等团队。在事件发生期间，这支核心团队每三个小时开一次会，直至所有站点都完成恢复、验证并返还给客户。

为管理恢复进度，我们创建了一个新的 Jira 项目 SITE 和一个工作流程，用于跨多个团队（工程、项目管理、支持等）逐个站点跟踪恢复情况。这种方法让所有团队都能轻松识别和跟踪在单个站点恢复过程中遇到的问题。

事件发生期间，我们于世界标准时间 4 月 8 日 03:30 对所有工程实施了代码冻结。这样我们能够专注于客户恢复，消除变化导致客户数据不一致的风险，将其他中断风险降至最低，并减少因无关变化而影响团队关注恢复的可能性。

事件时间线

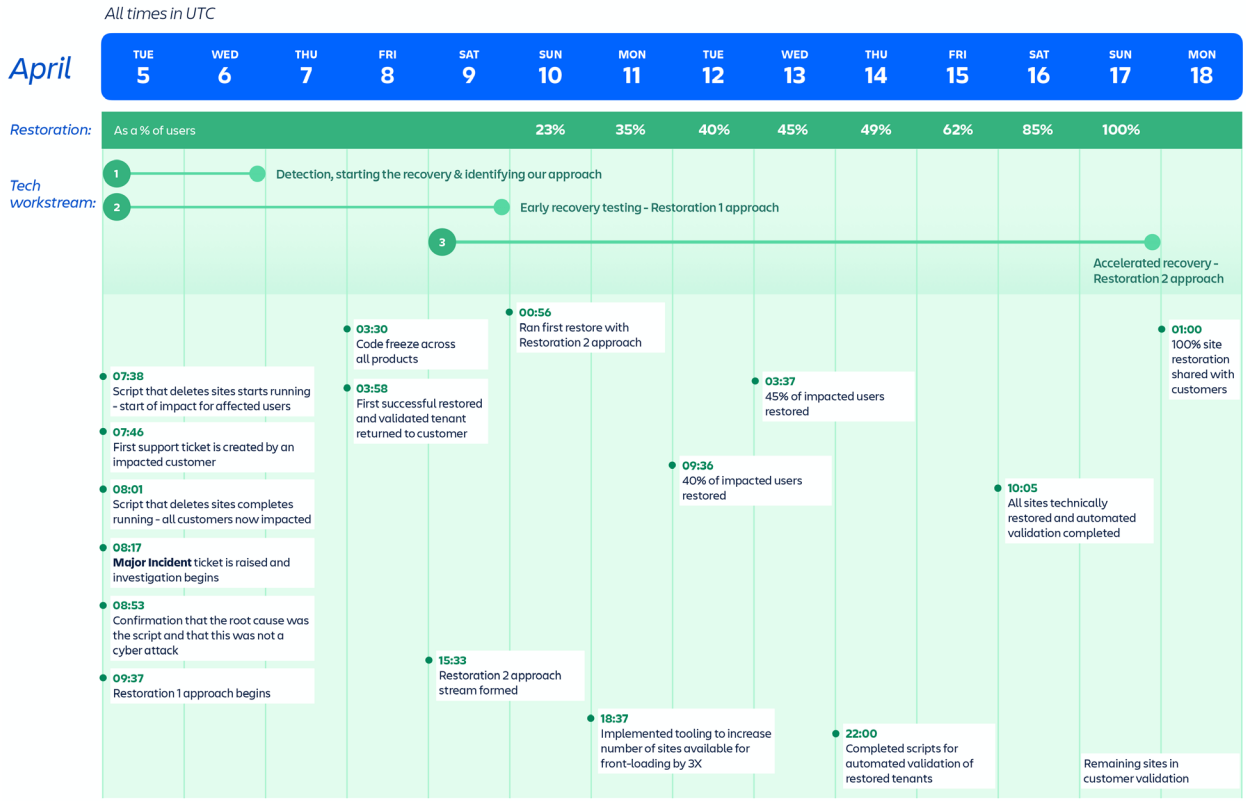


图 6：事件时间线和关键恢复节点。

恢复工作流的高级概述

恢复运行分为检测、早期恢复和加速这三个主要工作流。尽管下文针对每个工作流进行了单独介绍，但在恢复期间，所有工作流是并行发生的。

工作流 1：检测、开始恢复并确定我们的方式

时间戳：第 1-2 天（4 月 5 日至 6 日）

世界标准时间 4 月 5 日 08:53，我们发现 Insight 应用脚本造成站点被删除。我们确认这一情况不是由内部恶意行为或网络攻击造成。并立即召集相关产品和平台基础架构团队共同解决这一事件。

事件开始时，我们意识到：

- 恢复数百个被删除的站点涉及非常复杂的多步骤流程（详见上文架构部分），需要很多团队奋战多天才能顺利完成。
- 对我们来说，恢复单个站点没有问题，但是要恢复大批量的站点，我们还不具备相关能力和流程。

因此，我们需要大规模实现恢复过程并行化和自动化，以帮助受影响客户尽快恢复对 Atlassian 产品的访问权限。

工作流 1 需要大量开发团队执行以下活动：

- 为管道中的成批站点确定并执行恢复步骤。
- 编写并改进自动化功能，以便团队批量执行大量站点的恢复步骤。

工作流 2：早期恢复和“恢复 1”方法

时间戳：第 1-4 天（4 月 5 日至 9 日）

在脚本运行完成后一个小时，也就是世界标准时间 4 月 5 日 08:53，我们弄清了导致站点被删除的原因。我们还确定了恢复流程，该流程之前曾用于恢复少量站点的运行。但是，还未明确定义用于恢复如此大规模被删站点的恢复流程。

为了快速采取行动，该事件的早期阶段分给了以下两个工作组负责：

- 人工工作组负责验证所需步骤，并人工执行少量站点的恢复流程。
- 自动化工作组负责执行现有修复流程，并构建自动化，以便对更大批量的站点执行这些步骤。

“恢复 1”方法概述（见 *下方图 7*）：

- 这需要为每个被删除的站点创建新的站点，随后再为每个需要恢复数据的下游产品、服务和数据创建新站点。
- 新站点将获得新的 ID，比如 cloudId。这些 ID 都是不可变的，意味着很多系统会将这些 ID 嵌入到数据记录中。因此，如果这些 ID 发生改变，我们需要更新大量数据，这对于第三方生态系统应用来说尤为不妥。
- 修改新站点以复制被删除站点状态的各个步骤非常复杂，并且经常存在不可预见的依赖性。

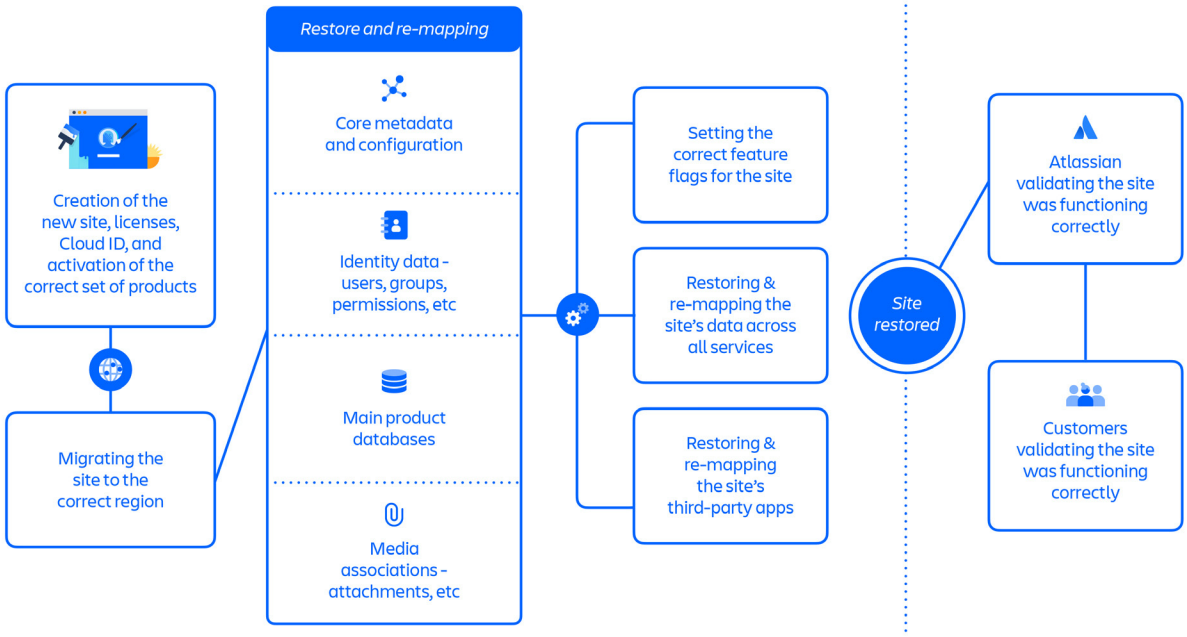


图 7：“恢复 1”方法中的关键步骤。

“恢复 1”方法包含约 70 个独立步骤，这些步骤在较高层次上汇总时，基本上会遵循一个连续流程：

- 创建新站点、许可证、Cloud ID 并激活正确的产品集
- 将站点迁移到正确的区域
- 恢复并重新映射站点的核心元数据和配置
- 恢复并重新映射站点的身份数据 – 用户、群组、权限等
- 恢复站点的主要产品数据库
- 恢复并重新映射站点的媒体关联 – 附件等

- 为站点设置正确的功能标记
- 跨所有服务恢复并重新映射站点数据
- 恢复并重新映射站点的第三方应用
- Atlassian 验证站点能否正常运行
- 客户验证站点能否正常运行

优化后，“恢复 1”方法需要约 48 小时恢复一批站点，自 4 月 5 日至 4 月 14 日，这种方法帮助 53% 的受影响用户恢复了 112 个站点。

工作流 3：加速恢复和“恢复 2”方法

时间戳：第 4-13 天（4 月 9 日至 17 日）
使用“恢复 1”方法，我们需要三个星期才能恢复所有客户。因此，4 月 9 日，我们提出了一种加快恢复所有站点的新方法，即“恢复 2”（见 *下方图 8*）。

“恢复 2”方法通过降低“恢复 1”方法的复杂性和依赖关系数量，提升了恢复步骤之间的并行性。

“恢复 2”方法涉及在所有相关系统中重新创建（或取消删除）与该特定站点相关的记录，首先是目录服务记录。这种新方法的一个关键因素是可以*沿用旧站点的 ID*。由此消除了之前的方法中用于将旧 ID 映射到新 ID 的一半以上的步骤，包括与每个站点的每个第三方应用供应商进行协调的步骤。

不过，从“恢复 1”方法改为“恢复 2”方法大幅增加了事件响应的开支：

- 在“恢复 1”方法中构建的很多自动化脚本和流程都要修改后才能用于“恢复 2”方法。
- 执行恢复的团队（包括事件协调员）必须在两种方法中管理并行的恢复批次，同时我们还要测试和验证“恢复 2”流程。
- 使用新方法意味着我们要在扩展“恢复 2”流程前对其进行测试和验证，这就需要重复之前对“恢复 1”执行的验证工作。

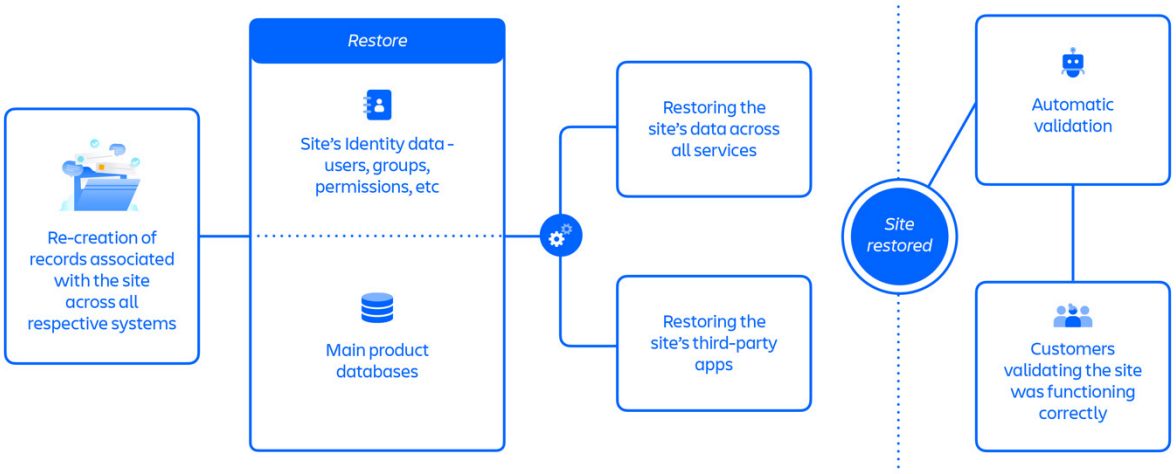


图 8：“恢复 2”方法中的关键步骤。

上图显示的是“恢复 2”方法，其中包含 30 多个步骤，这些步骤遵循基本并行的流程：

- 在所有对应系统中重新创建与站点关联的记录
- 恢复站点的身份数据 – 用户、群组、权限等
- 恢复站点的主要产品数据库
- 跨所有服务恢复站点数据
- 恢复站点的第三方应用
- 自动验证
- 客户验证站点能否正常运行

在加速恢复过程中，我们还采取了一些措施来预先加载和自动执行站点恢复，因为手动恢复不适合大批量恢复。恢复过程的顺序性意味着对于大型数据库恢复和用户基础/权限恢复，站点恢复可能会比较慢。我们完成的优化包括：

- 我们开发了所需工具和防护规则，以*预先加载*长期运行的步骤，如数据库恢复和身份同步，从而在其他恢复步骤之前完成这些步骤。
- 工程团队为各个步骤构建了自动化程序，从而可以安全执行大批恢复工作。
- 构建自动化程序是为了在所有恢复步骤完成后验证站点能否正常运行。

加速的“恢复 2”方法恢复一个站点需要约 12 个小时，自 4 月 14 日至 17 日，这种方法帮助 47% 的受影响客户恢复了 771 个站点。

恢复已删除的站点时，最大限度减少数据丢失

我们的数据库将完整备份与增量备份相结合来实施备份，因此我们可以选择特定“时间点”，以便在备份保留期（30 天）内恢复我们的数据存储。对于此次事件中的大部分客户，我们确定了我们产品的主要数据存储，并决定使用被删站点前五分钟的恢复点作为安全同步点。我们将非主数据存储恢复到了同一点，或通过重放记录事件进行了恢复。对主存储使用固定的恢复点，以便在所有数据存储中实现数据一致性。

对于我们在事件响应初期恢复的 57 个客户，由于缺乏一致的政策和手动检索数据库备份快照，导致一些 Confluence 和 Insight 数据库被恢复到了站点删除前五分钟*之前*的时间点。这种不一致是在后恢复审计过程中发现的。此后，我们又恢复了剩余的数据，联系了受此影响的客户，并帮助他们应用变更以进一步恢复其数据。

总结：

- 此次事件中，我们达到了一小时的恢复点目标 (RPO)。
- 事件中最多丢失了站点删除前五分钟的数据。
- 少数客户的 Confluence 或 Insight 数据库被恢复到了站点删除前五分钟之前的位置，不过我们可以恢复数据，目前正在与客户合作恢复相应数据。

事件沟通

谈到事件沟通，其涵盖与客户、合作伙伴、媒体、行业分析师、投资者和规模更大的技术社区的接触点。

情况说明

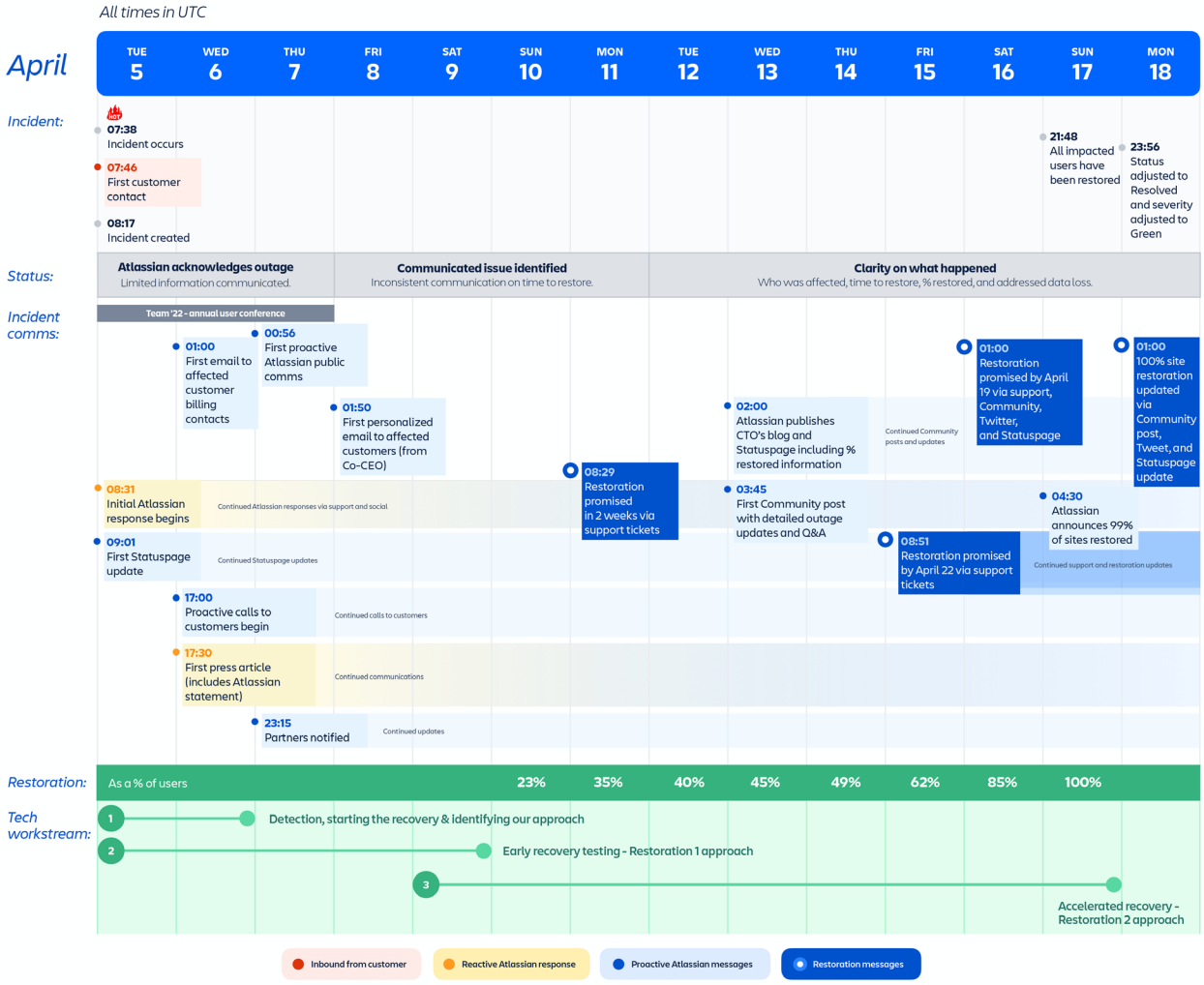


图 9：关键事件沟通节点时间线。

*时间戳：第 **1-3** 天（**4** 月 **5** 日至 **7** 日）*

早期响应

第一张支持请求单于世界标准时间 4 月 5 日 7:46 创建，Atlassian 支持人员于世界标准时间 8:31 之前做出了回应，确认了这一事件。在世界标准时间 9:03，发布了第一条 Statuspage 更新，告知客户我们正在调查这一事件。在世界标准时间 11:13，我们通过 Statuspage 确认我们已找到根本原因，并且我们在努力修复。截至世界标准时间 4 月 6 日 1:00，最初的客户请求单沟通表明中断源于维护脚本，我们预计数据丢失很少。Atlassian 于世界标准时间 4 月 6 日 17:30 发表声明，回应了媒体的询问。在世界标准时间 4 月 7 日 00:56，Atlassian 在推特上广泛发布了第一条外部消息，承认了这一事件。

*时间戳：第 **4-7** 天（**4** 月 **8** 日至 **11** 日）*

开始更广泛的个性化外展活动

世界标准时间 4 月 8 日 1:50，Atlassian 通过电子邮件向受影响客户发送了联合创始人兼联席首席执行官 Scott Farquhar 的道歉信。接下来的几天，我们努力恢复被删除的联系信息，并为所有尚未提交的受影响站点创建了支持请求单。然后，我们的支持团队继续通过每个受影响站点的相关支持请求单定期发送有关恢复工作的最新消息。

*时间戳：第 **8-14** 天（**4** 月 **12** 日至 **18** 日）*

事件情况更清晰，并完成了恢复

4 月 12 日，Atlassian 发布了来自首席技术官 Sri Viswanath 的更新，提供了更多技术细节，包括情况说明、受影响群体、有没有数据丢失以及我们的恢复进度，并说明了可能需要两周时间来完全恢复所有站点。该博客还另外附有一份 Sri 撰写的新闻声明。我们在工程主管 Stephen Deasy 主动发表的第一篇 Atlassian 社区文章中提到了 Sri 的博客，这篇博客后来成了更多更新和更广泛公众问答的专门渠道。4 月 18 日，该贴更新宣布所有受影响的客户站点均已全面恢复。

我们为什么没有早点公开回应？

1. 我们优先通过 Statuspage、电子邮件、支持请求单以及一对一互动直接与受影响客户进行了沟通。不过，有很多客户我们联系不到，因为随着客户站点被删除，我们也失去了他们的联系方式。我们应该更早发布更广泛的公告，向受影响客户的终端用户披露我们的事件响应和解决时间线。

2. 虽然我们很快就知道了事件的起因，但此次事件的架构复杂性和特殊性导致我们无法快速确定范围并准确预知问题的解决时间。我们不应等到全面了解所有情况后在做沟通，而应尽早披露我们当时已知和未知的情况，提供一般性恢复预估（可以是方向性预估），并明确说明我们预计什么时候可以全面了解所有情况，以便客户围绕事件更好地制定计划。这对系统管理员和技术联系人来说尤为重要，因其处在管理组织内部利益相关者和用户的前线。

支持体验和客户外联

如前所述，删除客户站点的脚本还从我们的生产环境中删除了关键客户 ID 和联系信息（例如，云 URL、站点系统管理员联系人)。这一点需要格外注意，因为我们的核心系统（例如支持、许可、计费）都利用了云 URL 和站点系统管理员联系人的存在性作为安全、路由和优先级的主要标识。一旦丢失这些标识，我们便无法第一时间系统性的识别客户并与之互动了。

我们会为受影响客户提供哪些支持？

首先，大部分受影响客户无法通过正常的在线联络表联系我们的支持团队。此表在设计上要求用户使用其 Atlassian ID 登录并提供有效的云 URL 才能使用。没有有效的 URL，用户无法提交技术支持请求单。在业务正常运行时，此项验证是为保障站点安全和请求单分流专门设立的。然而，对于受中断影响的客户，此项要求造成了意想不到的后果，导致他们无法提交高优先级的站点支持请求单。

其次，该事件还造成站点系统管理员数据被删除，继而导致我们难以主动与受影响的客户联系。在事件发生的最初几天，我们主动联系了 Atlassian 上登记的受影响客户的账单和技术联系人。然而，我们很快发现，很多受影响客户的账单和技术联系人信息已经过时。没有每个站点的系统管理员信息，我们就无法获得完整的活跃联系人和获批联系人列表。

我们是怎样应对的

我们的支持团队有三个同等重要的优先事项，即在事故发生的前几天加快站点恢复，并修复中断的沟通渠道。

第一，获得可靠且经过验证的客户联系人名单。我们的工程团队在努力恢复客户站点时，我们的客户团队会专注于经过验证的联系信息。我们会利用所掌握的每种机制（账单系统、过往支持请求单、其他安全的用户备份、直接的客户外联等）重新制定我们的联系名单，目的是要为每个受影响站点提供一张与事件相关的支持请求单，以简化直接外联和响应时间。

第二，重新建立针对此事件的工作流、队列和 SLA。云 ID 的删除以及无法正确认证用户影响了我们通过正常系统处理事件相关支持请求单的能力，请求单也无法正确显示到相关优先级和升级队列及控制面板中。我们迅速组建了一支跨职能团队（支持、产品、IT），以设计和添加额外的逻辑、SLA、工作流状态和控制面板。因为这项工作必须在我们的生产系统中完成，所以花了几天时间来全面开发、测试和部署。

第三，大规模扩展手动验证以加快站点恢复。随着工程团队在初始恢复中取得进展，很明显，我们需要全球支持团队协助通过人工测试和验证核实来加快站点恢复。当我们的工程团队加快数

据恢复速度后，这一验证过程将成为向客户提供恢复站点的关键路径。我们必须创建独立的标准操作程序 (SOP)、工作流、交接和人员名单流程，以动员 450 多名支持工程师进行验证核实，轮流提供全天候服务，加快将恢复的内容返还给客户。

尽管在第一周结束时，这些关键优先事项已经确定，但由于恢复过程非常复杂，事件处理的时间线也不够明确，我们提供*有意义*更新的能力受到了限制。我们应该早点承认我们无法提供确切的站点恢复日期，并尽早进行面对面讨论，以便我们的客户制定相应计划。

我们将如何改进？

我们立即阻止了批量站点删除，待适当更改完成后再开启此项功能。

在我们处理此次事件并重新评估我们的内部流程时，要认识到不要人为造成事故。相反，系统要允许犯错。本节总结了此次事件的成因。我们还讨论了为加快修复这些弱点和问题可采取的计划。

教训 1：应在所有系统设置通用“软删除”

总的来说，事件中的这类删除应该被禁止，或受到多层保护，以避免出错。我们采取的主要改进是在全球范围内防止跳过软删除程序直接删除客户数据和元数据。

a) 数据删除只能通过软删除进行

应禁止删除整个站点，软删除应要求多级保护以防止出错。我们将实施“软删除”策略，防止外部脚本或系统在生产环境中删除客户数据。我们的“软删除”策略支持充分的数据保留，以便快速、安全地执行数据恢复。只有在保留期过期后，才会从生产环境中删除数据。

行动：

- **在调配工作流和所有相关数据存储中实施“软删除”：**此外，租户平台团队还将验证数据删除是否只在停用后发生，以及该领域是否还有其他保护措施。从长远来看，租户平台将在进一步开发正确的租户数据状态管理方面发挥主导作用。

b) 软删除应具有标准化且经过验证的审核流程

软删除操作为高风险操作。因此，我们应设置标准化或自动化审核流程，其中包含明确的回滚和测试程序，以应对这些操作。

行动：

- **强制分阶段推出软删除操作：**所有需要删除的新操作首先都要在我们自己的站点内进行测试，以验证我们的方法和自动化程序。我们完成验证后，将会逐步引导客户完成同样的流程，并继续测试是否存在违规行为，然后才能将自动化应用于整个选定的用户群。

- **软删除操作必须具有经过测试的回滚计划：**在执行任何数据软删除活动前，必须对要删除的数据执行恢复测试，然后才能在生产中运行，并且要具有经过测试的回滚计划。

教训 2：作为灾难恢复计划的一部分，为更多客户的多站点、多产品删除事件提供自动恢复支持

Atlassian 数据管理详细描述了我们的数据管理流程。为确保高可用性，我们在多个 AWS 可用区域 (AZ) 调配和维护了一个同步备用副本。可用区域故障转移为自动化程序，通常可在 60-120 秒内完成。我们会定期处理数据中心中断和其他常见中断问题，不会对客户造成影响。

我们还会维护不可变备份，以抵御数据损坏事件，从而能够恢复到过去的时间点。
备份保留 30 天，Atlassian 会持续测试和审核存储备份以便恢复。必要时，我们可以帮助所有客户恢复到新的环境。

利用这些备份，我们可定期帮助意外删除自有数据的个别客户或小规模客户组回滚到过去的时间点。然而，站点级删除没有可快速自动化的运行手册，此次事件的规模需要以协调的方式对所有产品和服务进行工具化和自动化。

我们目前尚未实现自动化的是，在不影响其他客户的情况下，将大量客户的子设备恢复到现有（和目前正在使用的）环境。

在我们的云环境中，每个数据存储都包含来自多个客户的数据。由于在此次事件中被删除的数据只是其他客户持续使用的数据存储的一部分，所以我们必须从备份中手动提取和恢复相应片段。每个客户站点的恢复都涉及一个漫长而复杂的过程，恢复站点时需进行内部验证和最终客户验证。

行动：

- **面向更多客户加速多产品，多站点恢复：**灾难恢复 (DR) 计划满足我们现行的一小时 RPO 标准。我们将利用此次事件中的自动化程序和教训来加速灾难恢复计划，确保遇到此等规模事件也能满足我们在政策中明确的 RTO 标准。
- **将此案例中的验证程序自动化并添加到灾难恢复测试中：**我们将定期进行灾难恢复练习，包括为大量站点恢复所有产品。这些灾难恢复测试将验证随着我们架构的发展以及新边缘案例的出现，我们的运行手册有没有对应更新。我们将不断改进我们的恢复方法，自动化更多的恢复流程，并缩短恢复时间。

教训 3：改进大规模事件的事件管理流程

我们的事件管理计划非常适合管理多年来发生过的大小事件。我们经常会针对一些规模较小、持续时间较短的事件进行应急响应模拟训练，这类事件通常涉及的人员和团队较少。

然而，此次事件在高峰期要求数百名工程师和客户支持员工同时工作，以恢复客户站点。这类事件的深度、广度和持续时间是我们的事件管理计划和团队始料未及的（见 *下方图 10*）。

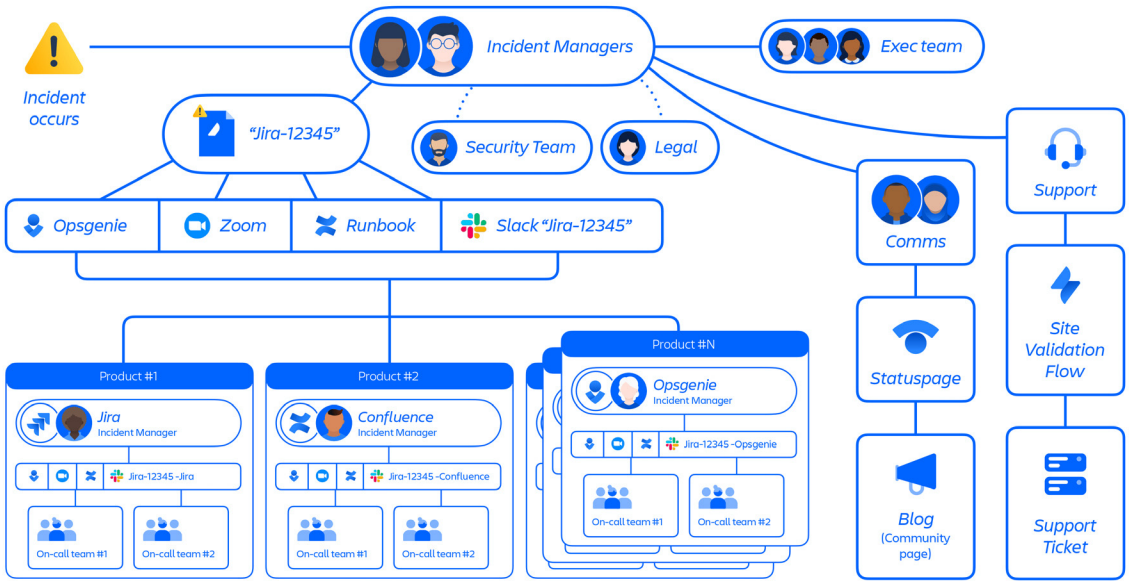


图 10：大规模事件管理流程概述。

我们将会优化大规模事件管理流程的定义并加强训练

我们有针对产品级事件的行动手册，但都不适用于要求全公司数百名员工同时处理的此等规模事件。在事件管理工具中，我们有创建 Slack、Zoom 和 Confluence doc 通信流的自动化程序，但不能创建适用于大规模事件以隔离恢复流的子通信流。

行动：

- **为大规模事件定义行动手册和工具，并进行模拟训练：**定义并记录可能应视为大规模事件的类型以及这类事件所需的响应级别。概述关键协调步骤并构建工具来帮助事件经理和其他业务职能部门简化响应并启动恢复。团队中的事件经理将定期安排模拟训练和培训，完善工具和文件，以不断改进。

教训 4：改进我们的沟通流程

a) 我们删除了重要的客户 ID，导致我们难以与受影响客户沟通并采取行动

删除客户站点的脚本同时还从我们的生产环境中删除了重要的客户 ID（例如云 URL、站点系统管理员联系人)。由此导致，(1) 客户无法通过我们的正常支持渠道提交技术支持请求单；(2) 我们花了数天时间才获得了受中断影响的关键客户联系人（例如站点系统管理员）的可靠名单，以主动发起互动；(3) 因此次事件的特殊性导致支持工作流、SLA、控制面板和升级程序无法正常运行。

此次中断期间，客户升级还是通过多渠道（电子邮件、电话、首席执行官请求单、领英和其他社交渠道，以及支持请求单）进行的。我们的客户团队使用了不同的工具和流程，由此减缓了我们的响应速度，并加大了全面跟踪和报告客户升级的难度。

b) 我们没有足够详尽的事件沟通行动手册来应对如此复杂的事件

我们没有概述各项原则及角色和职责的事件沟通行动手册，无法足够快速地动员统一的跨职能事件沟通团队。我们没有通过多渠道，尤其是社交媒体渠道，快速而一致地承认这一事件。我们本应围绕此次中断提供更广泛的公告，并重复关键信息，即没有数据丢失，也不存在网络攻击。

行动：

- **改进关键联系人备份：**在产品实例以外备份授权客户联系人信息。
- **改造支持工具：**为没有有效站点 URL 或 Atlassian ID 的客户建立机制，以便与我们的技术支持团队直接联系。
- **客户升级系统和流程：**投资建设基于账户的统一升级系统和工作流，将多个工作对象（请求单、任务等）存储在单个客户帐户对象之下，以改善所有客户团队的协调与可见性。
- **加快推进全天候升级管理覆盖率：**根据升级管理职能的全球足迹拓展计划执行，让每个主要地理区域的指定人员和支持角色都能提供一致的全天候覆盖服务，以协助相关产品和销售主题专家和领导的工作。
- **获得新的经验时，及时更新我们的事件沟通行动手册，并定期回顾：**回顾行动手册，以便在内部定义明确的角色和沟通路径。使用 DACI 事件框架，并为各个角色实施全天候备份，以应对生病、休假或其他不可预知的情况。执行季度审计，以随时验证就绪情况。

操作（续）

按照事件通信模板进行所有通信：提供情况说明、受影响群体、恢复时间表、站点恢复百分比、数据丢失预估、相关置信水平，以及关于支持人员联系方式的明确说明。

结束语

随着中断得到解决，客户得到全面恢复，我们的工作也还在继续。现阶段，
我们正在实施上述变革，以改进我们的流程，提高我们的灵活性，并防止类似事件再次发生。

Atlassian 是一个学习型组织，我们的团队必定已经从此次经历中吸取了刻骨铭心的教训。
我们正在将这些经验教训应用到实践中，确保对我们的业务产生持久性的改变。最后，
经过此次事件后，我们将变得更强大，并将为您提供更好的服务。

我们希望从此次事件中吸取的教训，对其他勤勤恳恳为客户提供可靠服务的团队也能有所帮助。

最后，我要感谢本文的各位读者，感谢与我们一同学习的人员以及加入 Atlassian 社区和团队的人员。

- 首席技术官 Sri Viswanath