ATLASSIAN

事件回顧檢討

2022 年 4 月停機事件

此翻譯僅供參閱。如果譯文有任何不清楚或矛盾之處請以英文原文為準。

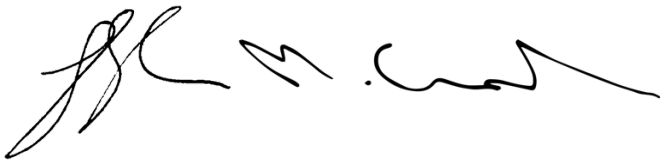
來自共同創辦人與共同執行長的一封信

我們要承認本月稍早因為停機，造成客戶服務中斷。我們了解 Atlassian 產品對您的業務至關重要，我們十分看重這份責任。針對這次事件，我們責無旁貸。我們會全權負責。針對受影響的客戶，我們也正積極贏回您的信任。

Atlassian 的核心價值觀之一是「開放的公司，絕無虛言」。因為這樣的價值觀是我們生活的一部分，我們會公開討論事件，並將事件視為學習機會。我們為客戶、Atlassian 社群和廣大的技術界發佈了這次的事件回顧檢討，並對 Atlassian 的事件管理流程感到驕傲。其中強調了不怪罪員工的公司文化，而是著重在找出技術系統和流程的改善方法，這對於提供大規模又值得信賴的服務來說至關緊要。雖然我們會盡力避免任何類型的事件，但也接受危機就是轉機這樣的想法。

請您放心，Atlassian 的雲端平台讓我們能滿足超過 20 萬名使用者，不同規模、不同行業的需求。在此次事件發生之前，我們的雲端持續提供了 99.9% 的正常運行時間以及超越正常運行時間的 SLA。我們長期在平台、幾個集中式平台功能進行投資，並備有可擴展的基礎架構與穩定的安全性增強功能。

我們向客戶與合作夥伴致上誠摯的感謝，謝謝您持續信任 Atlassian 並與我們合作。我們希望這份文件的詳細內容與行動可充分顯示，Atlassian 會繼續提供世界級的雲端平台以及強大的項目組合，藉此滿足每一個團隊的需求。



-Scott 與 Mike

執行摘要

2022 年 4 月 5 日 (星期二) 7:38 UTC (下列時間皆為世界協調時間) 開始，有 775 位 Atlassian 客戶無法存取 Atlassian 產品。部分客戶的停機時間長達 14 天，第一批客戶於 4 月 8 日恢復使用，其餘所有客戶網站皆於 4 月 18 日之前逐漸恢復。

這起事件並非是由網路攻擊造成的，也不是因為他人未經授權即存取客戶的資料。Atlassian 有一項全方位的資料管理計畫，其中包含已發佈的 SLA 以及超過 SLA 的歷史。

雖然此次事件為重大事件，但沒有任何客戶遺失超過 5 分鐘的資料。此外，在復原活動期間，有超過 99.6% 的客戶與使用者仍能持續使用 Atlassian 雲端產品，並未受到影響。

> 在整份文件中，我們將此事件中網站遭到刪除的客戶稱為「受影響」的客戶。
>
> 此事件回顧檢討提供了事件的確切詳細資料，概述了我們為恢復所採取的步驟，並描述了我們將如何防止此類情況在未來發生。我們在本節中提供了事件的整體摘要，並在文件的其餘部分中進一步詳細介紹。

事件始末

2021 年，我們完成了 Jira Service Management 和 Jira Software 的 Atlassian 獨立應用程式收購與整合，應用程式名為「Insight – Asset Management」。此獨立應用程式的功能透過原生形式整合到 Jira Service Management 中，不再適用於 Jira Software。因此，我們需要刪除客戶網站上安裝的獨立舊版應用程式。工程團隊使用現有的指令碼和流程，刪除此獨立應用程式的執行個體。然而，發生了兩個問題：

- **溝通落差。**請求刪除的團隊與執行刪除的團隊之間有溝通落差。請求刪除的團隊*沒有提供*標記刪除的目標應用程式 ID，*反而提供了*應用程式所在的整個雲端網站 ID。

- **系統警告不足。**用於執行刪除的 API 接受了網站和應用程式的識別碼，並假定輸入正確。因此若是傳送網站 ID，網站便會遭到刪除；若是傳送應用程式 ID，應用程式便會遭到刪除。沒有警告訊號確認要求的 (網站或應用程式) 刪除類型。

執行的指令碼遵循標準同行審查流程。標準同行審查流程著重在呼叫哪個端點以及呼叫方式。流程並未交叉檢查提供的雲端網站 ID，也並未驗證 ID 是指 Insight 應用程式還是整個網站。問題就在於指令碼包含了客戶整個網站的 ID。因此造成 2022 年 4 月 5 日星期二 07:38 到 08:01 這段時間內，有 883 個網站 (775 位客戶) 立即遭到刪除。*請參閱「事件始末」*

我們如何回應？

於 4 月 5 日 08:17 確認事件發生後，我們便啟動重大事件管理流程，並建立交互功能事件管理團隊。全球事件回應團隊在事件發生期間不眠不休持續工作，直到復原所有網站後，驗證網站並交還給客戶。此外，事件管理負責人每三小時舉行一次會議，協調工作流程。

之前我們便意識到要同時復原數百位客戶的多種產品，會面臨諸多挑戰。

事件發生時，我們便確認過有哪些網站受到影響。當下的首要任務就是與網站的核准擁有者溝通，通知他們發生停機事件。

不過，有某些客戶的連絡資訊已經遭到刪除。代表客戶無法像以往一樣送出支援工單，而我們也無法立刻存取重要的客戶連絡人。*如需更多詳細資料，請參閱「復原工作流程的整體概觀」*

我們採取了什麼行動，以防止類似事件再次發生？

我們立即採取行動，並致力做出改變，以避免未來發生類似情形。以下是我們已經做出 (或將會做出) 重大改變的四個特定領域：

1. **在所有系統建立通用的「軟刪除」功能。**整體而言，本事件類型的刪除方式應受到禁止，或者具有多層保護以避免發生錯誤，其中包含階段性推出及測試「軟刪除」復原計畫。我們會全面避免尚未經過軟刪除流程的客戶資料和中繼資料遭到刪除。

2. **加速災難復原 (DR) 計畫，在遭遇多網站、多產品的刪除事件時，為更多客戶提供自動化復原。**我們將使用自動化功能，以及運用此次事件中學習到的經驗加速災難復原計畫，以符合我們在原則中，針對此事件規模所設立的復原時間目標 (RTO)。我們將定期進行災害復原演練，了解如何復原大量網站的所有產品。

3. **修訂大規模事件管理流程。**我們會改善大規模事件的標準作業程序 (SOP)，並透過模擬此事件的規模進行演練。我們也將更新訓練方式與工具，以利大量團隊協同工作。

4. **建立大規模事件通訊教戰守則**。我們會透過多重管道儘快確認事件經過，也將在數小時內發佈與事件相關的公開通訊資料。為了能確實連絡受到影響的客戶，未來我們也會改善重要連絡人的資料備份，並改良支援工具，讓沒有有效 URL 或 Atlassian ID 的客戶能夠直接與技術支援團隊聯繫。

如需完整的交辦事項列表，請詳見下方事件回顧檢討內容。*請參閱「我們將如何改進」*

目錄

Atlassian 雲端架構概述

先了解 Atlassian 產品、服務和基礎架構的部署架構，將有助於了解本文中討論的事件發生原因。

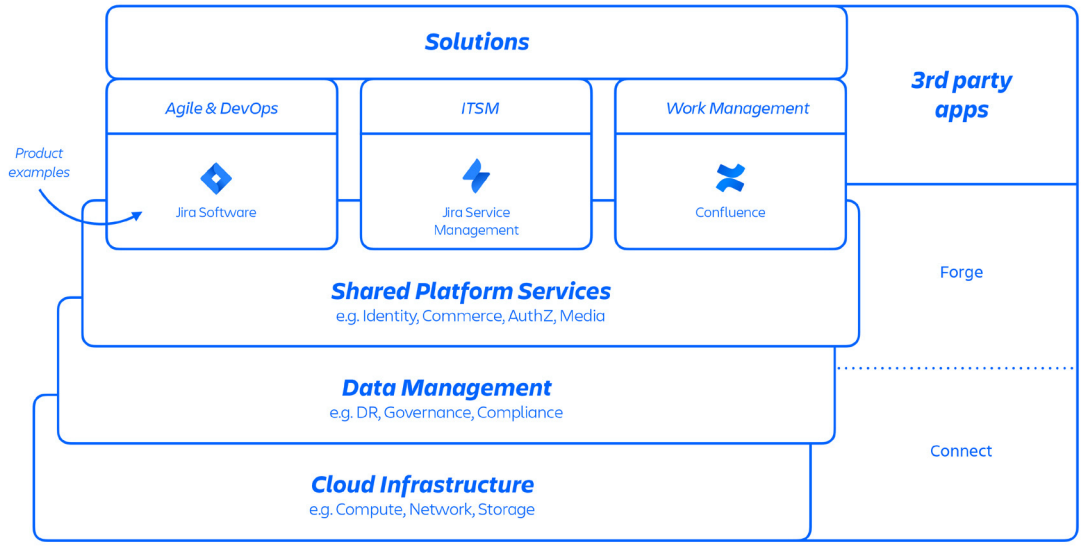
Atlassian 雲端裝載架構

Atlassian 使用 Amazon Web Services (AWS) 作為雲端服務供應商，並使用 AWS 在全球多個地區設置的高可用性資料中心設施。每個 AWS 區域都位於獨立的地理位置，包含多個獨立運作且置放在不同位置的資料中心，稱為可用區域 (AZ)。

我們使用 AWS 的計算、儲存空間、網路與資料服務來建立產品和平台元件。因此我們能夠利用 AWS 提供的冗餘功能，例如可用區域與區域。

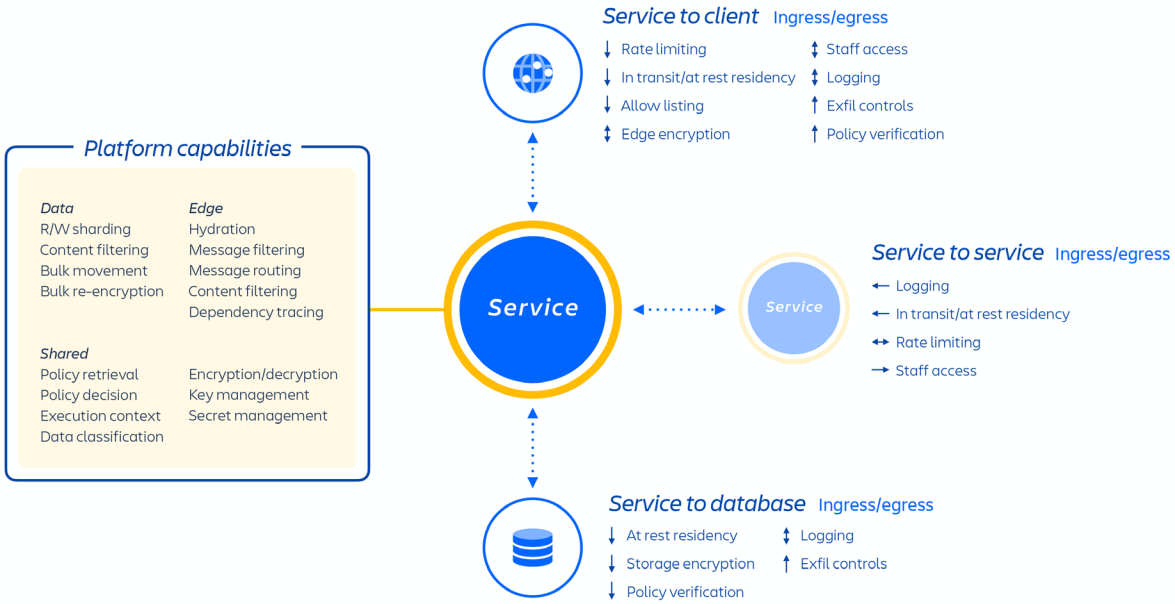
分散式服務架構

藉由此 AWS 架構，我們提供許多平台與產品服務，也運用在我們的解決方案中。其中包含多個平台功能，能與眾多 Atlassian 產品共享及使用，例如 Media、Identity、Commerce 以及 Editor 等體驗。除此之外還有產品特定功能，例如 Jira Issue 服務以及 Confluence Analytics。



圖表 1：Atlassian 平台架構。

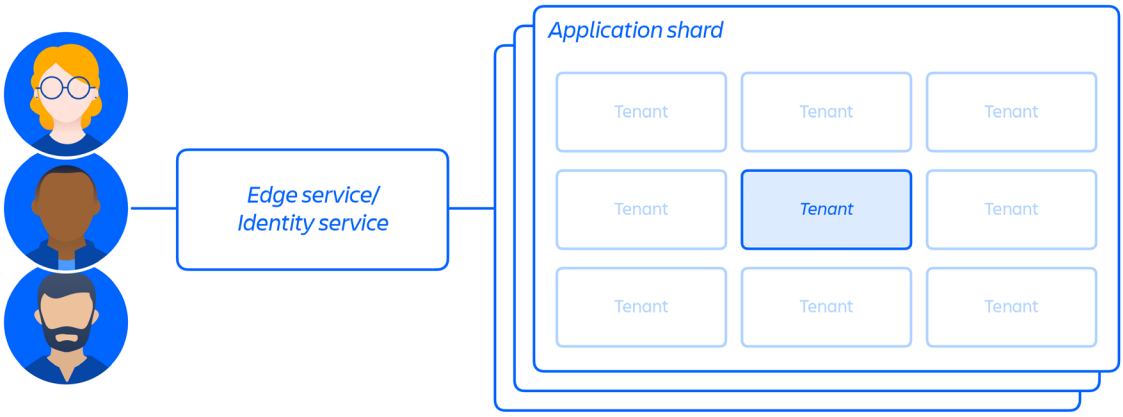
Atlassian 開發人員透過內部開發的平台即服務 PaaS (也稱為 Micros)，佈建上述服務，能自動協調共用服務部署、基礎架構、資料儲存以及管理功能，包含安全性與合規性控制要求 *(請參閱上方圖表 1)*。Atlassian 產品通常是由多個使用 Micros 部署在 AWS 上的「容器化」服務組成的。Atlassian 產品使用核心平台功能 *(請參閱下方圖表 2)*，涵蓋要求路由、二進位物件儲存、驗證/授權、交易性使用者生成內容 (UGC)、實體關係儲存、資料湖泊、一般記錄、要求追蹤、可檢視性以及分析服務。這些微服務是使用平台層級標準化、經核准的技術堆疊所建立：



圖表 2：Atlassian 微服務概覽。

多租戶架構

除了雲端基礎架構之外，我們也建立並運作多租戶微服務架構，以及支援 Atlassian 產品的共用平台。在多租戶架構中，單一服務即可為多位客戶提供服務，其中包含執行雲端產品所需的資料庫和運算執行個體。每個分區 (本質上為容器 - 請參閱 *下方圖表 3*) 含有多位租戶的資料，不過每個租戶的資料是獨立的，其他租戶無法存取。



圖表 3：我們如何在多租戶架構中儲存資料。

租戶佈建與生命週期

新客戶佈建後，有一系列的活動會觸發分散式服務協調流程和資料儲存佈建。這些活動通常會對應到生命週期的七個步驟之一：

1. 商務系統會立即更新最新的客戶中繼資料與存取控制資訊。接著藉由一系列租戶與產品活動，佈建協調系統便會核對「佈建資源狀態」與授權狀態。

租戶事件

這些事件會為全體租戶帶來一體兩面的影響：

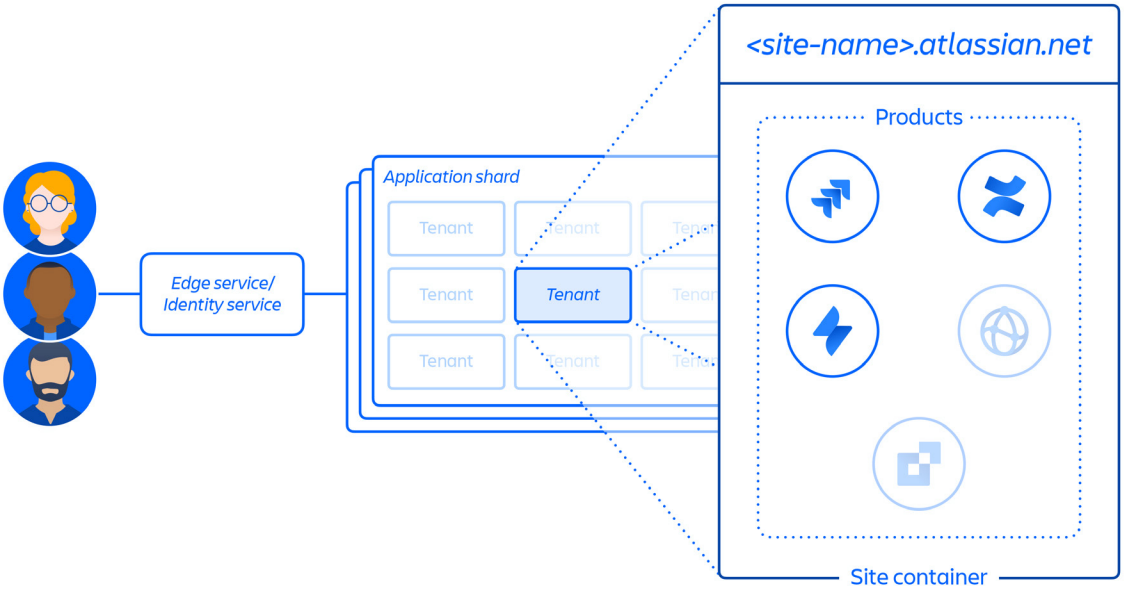
- 創造：客戶建立了租戶並用於全新的網站
- 破壞：整個租戶遭到刪除

產品事件

- 啟用：啟用授權產品或第三方應用程式後
- 停用：停用特定產品或應用程式後
- 暫停：暫停給定的現有產品後，同時禁止存取其所有的給定網站
- 取消暫停：在取消暫停給定的現有產品後，同時允許存取其所有的網站

授權更新：包含有關給定產品的授權基座數量資訊及其狀態 (活躍/非活躍)

2. 建立客戶網站並為客戶啟用正確的產品集。網站的概念，是作為一個裝有許多產品的容器，並且授權給特定客戶 (例如 Confluence 以及 Jira Software 的<網站名稱>.atlassian.net)。(請參閱 *下方圖表 4*)若要了解此報告的內容，這點十分重要，因為這次事件中遭到刪除的便是網站容器，而整份文件也一直在討論網站的概念。



圖表 4：網站容器概覽。

3. 在指定區域的客戶網站上佈建產品。

   佈建產品後，大部分的內容將裝載在使用者存取產品的位置附近。為了最佳化產品效能，我們不會限制全域裝載的資料移動。若有需要，我們可能會在區域之間移動資料。

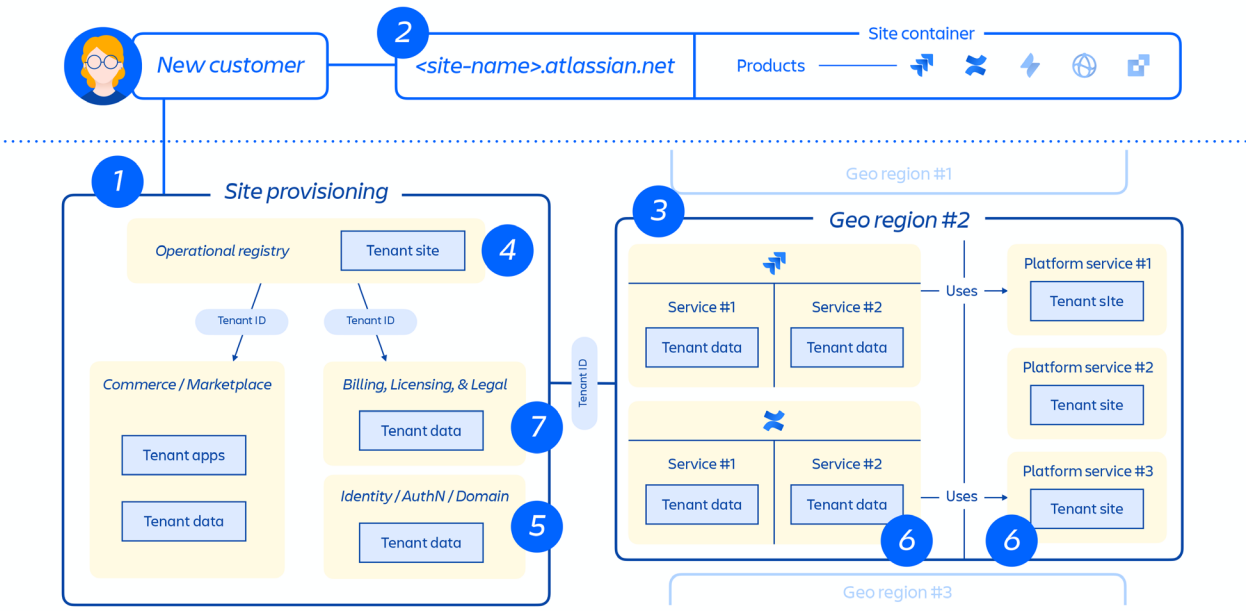
   針對某些產品，我們也提供資料落地。資料落地讓客戶能選擇產品資料是要全域分散，還是儲存在我們定義的其中一個地理位置。

4. 建立及儲存客戶網站以及產品核心中繼資料與配置。

5. 建立及儲存網站以及產品身份識別資料，例如使用者、群組、權限等。

6 在網站內佈建產品資料庫，例如：Jira 產品系列、Confluence、Compass 與 Atlas。

7 佈建產品授權的應用程式。



圖表 5：概攬客戶網站如何在分散式架構中佈建。

上方的圖表 5 展示了客戶的網站不僅僅部署在單一資料庫或儲存區，也部署在我們的分散式架構中。其中包含多個實體位置與邏輯位置，用來儲存中繼資料、配置資料、產品資料、平台資料以及其他相關的網站資訊。

災難復原計畫

我們的災難復原 (DR) 計畫致力於針對基礎架構失敗提供復原，以及從備份中復原服務儲存空間。若要了解災難復原計畫，有以下兩個重要概念：

- **復原時間目標 (RTO)**：災難發生期間，復原資料並交回給客戶的速度有多快？
- **復原點目標 (RPO)**：從備份復原後，復原的資料是多近的資料？自上次備份後會遺失多少資料？

在這次事件中，我們未達成復原時間目標，但達成了復原點目標。

復原

我們已經為基礎架構層級失敗做準備；舉例來說，整個資料庫、服務或 AWS 可用區域遺失。這項準備工作包含資料及服務複寫，橫跨多個可用區域並定期執行容錯移轉測試。

服務儲存空間可復原性

因勒索軟體、惡意執行者、軟體缺失或操作錯誤導致的服務儲存區資料損毀，我們也已做好復原準備。這項準備工作包含不可變備份以及服務儲存區的備份復原測試。我們能夠將任何單獨的資料儲存區復原到之前的時間點。

多網站、多產品自動復原性

事件發生時，我們無法選取大量客戶網站，並將所有互相連接的產品從備份復原到之前的時間點。

我們的功能一直專注於基礎架構、資料損毀、單一服務活動或單一網站刪除。過去我們不得不處理並測試這類錯誤。網站級別的刪除沒有能夠快速自動執行此事件規模的運行手冊，這需要以協調的方式跨所有產品和服務進行工具和自動化。

接下來的章節將更深入地介紹此複雜度，以及我們在 Atlassian 正持續進行的工作，藉此發展及最佳化我們的能力，以大規模維護本架構。

事件始末、時間表與復原

事件始末

2021 年，我們完成了 Jira Service Management 和 Jira Software 的 Atlassian 獨立應用程式整合，應用程式名為「Insight – Asset Management」。此獨立應用程式的功能透過原生形式整合到 Jira Service Management 中，不再適用於 Jira Software。因此，我們需要刪除客戶網站上安裝的獨立舊版應用程式。工程團隊使用現有的指令碼和流程，刪除此獨立應用程式的執行個體。

然而，發生了兩個重大問題：

- **溝通落差。**請求刪除的團隊與執行刪除的團隊之間有溝通落差。請求刪除的團隊沒有提供標記刪除的目標應用程式 ID，反而提供了應用程式所在的整個雲端網站 ID。
- **系統警告不足。**用於執行刪除的 API 接受了網站和應用程式的識別碼，並假定輸入正確。因此若是傳送網站 ID，網站便會遭到刪除；若是傳送應用程式 ID，應用程式便會遭到刪除。沒有警告訊號確認要求的 (網站或應用程式) 刪除類型。

執行的指令碼遵循標準同行審查流程。標準同行審查流程著重在呼叫哪個端點以及呼叫方式。流程並未交叉檢查提供的雲端網站 ID，也並未驗證 ID 是指應用程式還是整個網站。該指令碼已經在預備環境中按標準變更管理流程測試，不過由於預備環境中沒有 ID，因此指令碼並未偵測到 ID 輸入不正確。

在正式環境執行時，最初指令碼執行了 30 個網站。第一次正式環境執行成功，刪除了這 30 個網站的 Insight 應用程式，沒有其他負面影響。但是，這 30 個網站的 ID 是在溝通不良事件之前取得的，因此不包含正確的 Insight 應用程式 ID。

後續正式環境執行的指令碼包含了網站 ID，而非 Insight 應用程式 ID，並針對 883 個網站執行。指令碼於 4 月 5 日 07:38 開始執行，並於 08:01 執行完成。指令碼根據輸入列表依序刪除網站，因此指令碼在 07:38 開始執行不久後，第一個客戶的網站就遭到刪除。最後導致 883 個網站遭到立即刪除，且我們的工程團隊並未收到警告訊號。

這導致受影響的客戶無法使用下列 Atlassian 產品：Jira 產品系列、Confluence、Atlassian Access、Opsgenie 與 Statuspage。

一得知此事件發生，我們的團隊就開始努力為所有受影響的客戶進行復原。當時，我們估計受影響的網站數量約為 700 個 (共有 883 個網站受影響，但我們減掉了 Atlassian 擁有的網站數量)。在這 700 個網站中，大部分是免費、不活躍的帳戶，以及活躍使用者數量較少的小帳戶。基於上述原因，我們初步估計受影響的客戶約為 400 位。

我們現在對於資訊有更清楚的掌握，基於 Atlassian 官方客戶定義的完全透明度原則，我們想表示，總共有 775 位客戶受到停機影響。不過，大多數使用者都涵蓋在最初估計的 400 位客戶中。其中有一小部分的客戶的停機時間長達 14 天，第一組客戶於 4 月 8 日復原，其餘客戶則在 4 月 18 日全數復原。

我們如何協調

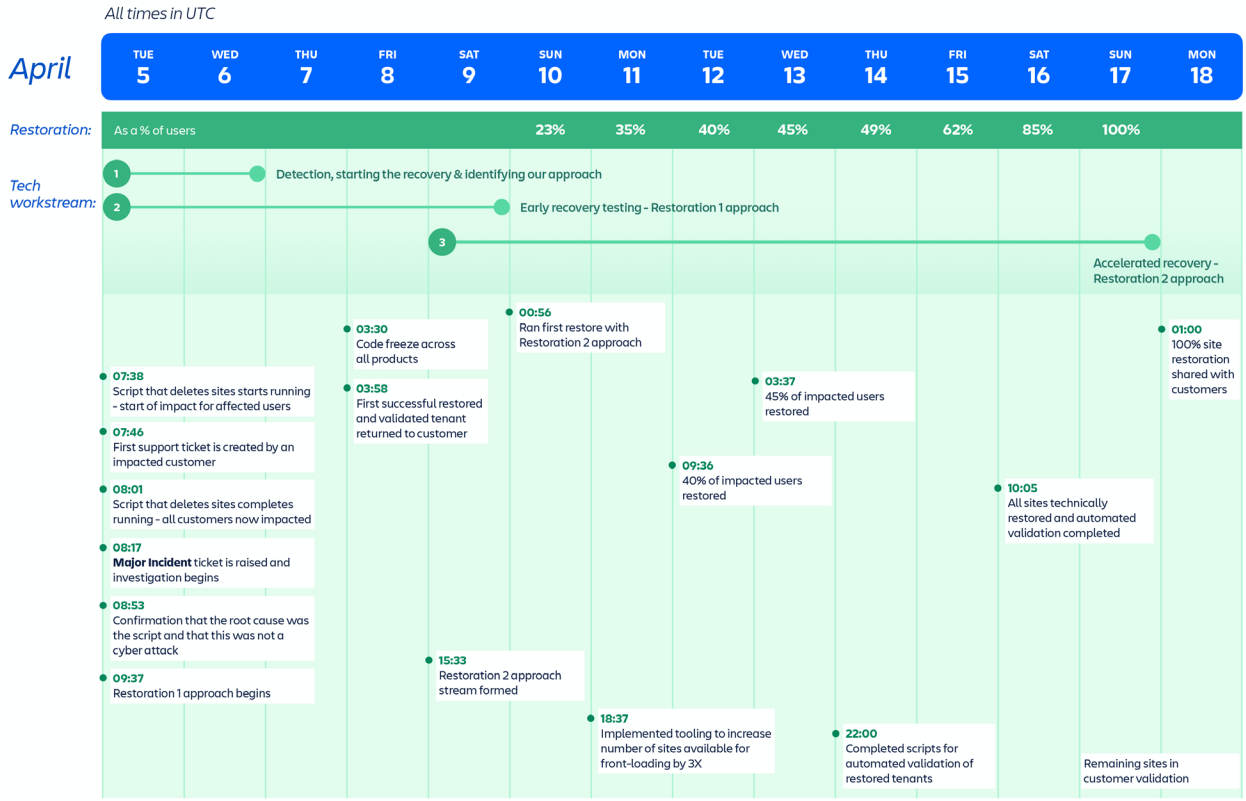
第一個支援工單由受影響的客戶於 4 月 5 日 07:46 時建立。由於這些網站是透過標準工作流程刪除的，所以我們的內部監控並未偵測到問題。我們在 08:17 時觸發重大事件管理流程，並建立交互職務事件管理團隊。在 7 分鐘內 (08:24) 時，便已將該事件向上呈報為「危急」狀態。我們的團隊在 08:53 時確認客戶的支援工單與指令碼執行有關。一意識到復原的複雜度，我們便在 12:38 時將該事件的嚴重性列為最高等級。

事件管理團隊由 Atlassian 不同團隊的成員組成。成員來自工程、客戶支援、專案管理、溝通等團隊。核心團隊在事件期間每三小時舉行一次會議，直到復原所有網站後，驗證網站並交還給客戶。

為了管理復原進度，我們建立一個新的 Jira 專案、網站與工作流程，以便追蹤 (工程、專案管理、支援等) 跨團隊的行動，逐一將網站復原。這項方法讓所有團隊能輕鬆識別任何網站復原的相關問題，並加以追蹤。

事件發生期間，我們於 4 月 8 日 03:30 時，對所有工程部門實施程式碼凍結。這讓我們能專注於復原工作，消除會導致客戶資訊不一致的變更風險，降低其他停機風險，並減少會使團隊分神的不相關變更。

事件時間表



圖表 6：事件的時間表和重要的復原里程碑。

復原工作流程的整體概觀

復原是以三個主要工作流程執行 - 偵測、早期復原與加速復原。雖然下列內容中，我們分別描述各個工作流程，但復原期間所有工作流程都是並行的。

工作流程 1：偵測，啟動復原以及確認方法

時間戳記：第 1-2 天 (4 月 5 日至 6 日)

我們在 4 月 5 日 08:53 時，知道是 Insight 應用程式指令碼導致網站遭到刪除。我們確認這並非是內部惡意行為或網路攻擊造成的。相關的產品與平台基礎架構團隊皆被安排處理此事件。

事件發生一開始時，我們便瞭解到：

- 復原數百個已刪除的網站過程複雜且涉及許多步驟，需要許多團隊及許多天才能完成 (過程在上方的架構章節有詳細描述)。
- 我們能夠復原單一網站，但尚未建立復原大批網站的功能與流程。

因此，我們需要大幅度的自動化與並行復原過程，讓受影響的客戶能儘快恢復他們 Atlassian 產品的存取權限。

工作流程 1 需要大量開發團隊參與下列活動：

- 為管道中的網站識別並執行復原步驟。
- 編寫並改進自動化，讓團隊在每一批次中能對更多網站執行復原步驟。

工作流程 2：早期復原與「復原方法 1」

時間戳記：第 1-4 天 (4 月 5 日至 9 日)

在 4 月 5 日 08:53 時 (也就是指令碼執行完成的一小時內)，我們已經瞭解導致網站遭到刪除的原因。我們也知道之前在正式作業環境用來復原少數網站的復原過程。但是，之前尚未明確定義這類規模的已刪除網站還原過程。

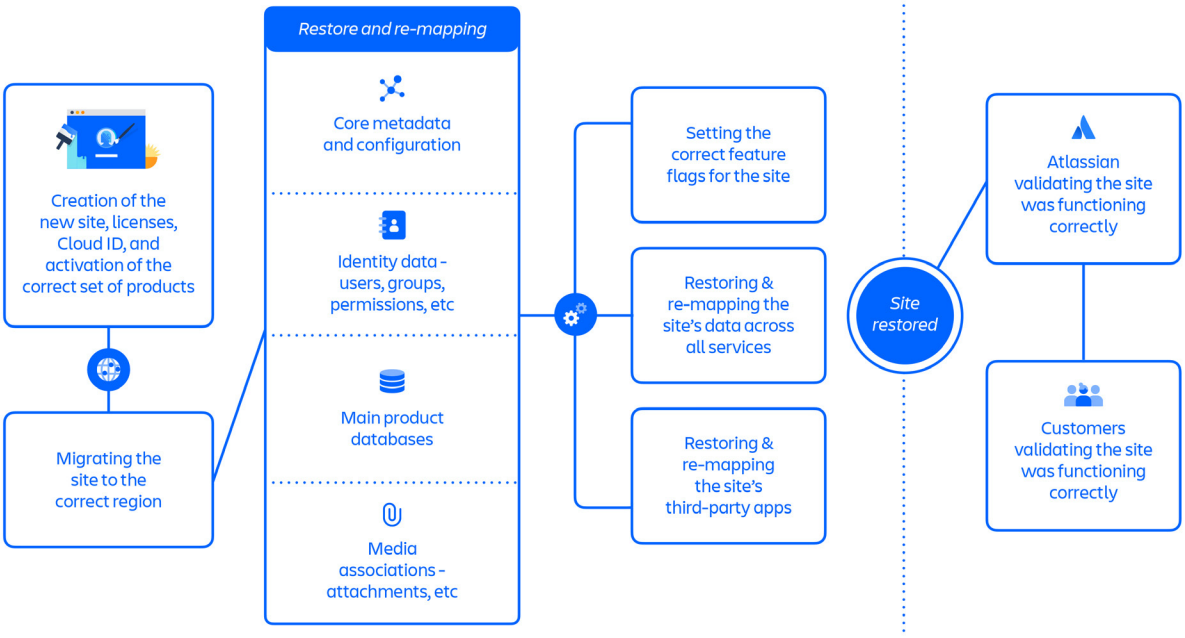
為了快速採取行動，事件早期階段便安排兩個工作組別：

- 手動工作組驗證了所需的步驟，並手動執行少量網站的復原流程。
- 自動化工作組則採用現有的復原流程並建立自動化，以便能在更大批次的網站上安全地執行復原步驟。

「復原方法 1」概覽 (請參閱 *下方圖表 7*)：

- 需要為每個已刪除的網站建立新網站，還有每個下游產品、服務、以及還原資料所需的資料儲存空間。
- 新建立的網站會有新的識別碼，例如 `cloudId`。這些識別碼是不可變的，這代表許多系統已經將這些識別碼嵌入到資料記錄中。因此如果變更識別碼，我們便需要更新大量的資料，這會造成第三方生態系統應用程式的問題。

- 如果修改新網站以複寫已刪除網站的狀態，會在復原步驟中建立十分複雜且通常無法預見的相依性。



圖表 7：復原方法 1 的重要步驟。

「復原方法 1」包含約 70 個單獨步驟。在高層級彙總時，這些步驟主要遵循以下各個流程：

- 建立新網站、授權、雲端 ID 以及啟用正確的產品集
- 將網站移轉到正確的區域
- 復原並重新對應網站的核心中繼資料和配置
- 復原並重新對應網站的身份識別資料 - 使用者、群組、權限等
- 復原網站的主要產品資料庫
- 復原並重新對應網站的媒體關聯 - 附件等
- 為網站設定正確的功能標記
- 復原並重新對應網站所有服務的資料
- 復原並重新對應網站的第三方應用程式
- Atlassian 驗證網站是否能正常運作
- 客戶驗證網站是否能正常運作

「復原方法 1」在最佳化後，需要約 48 小時才能復原一批次的網站，並在 4 月 5 日至 4 月 14 日期間用於復原 112 個網站中，53% 的受影響使用者。

工作流程 3：加速復原與「復原方法 2」

時間戳記：第 4-13 天 (4 月 9 日至 17 日)

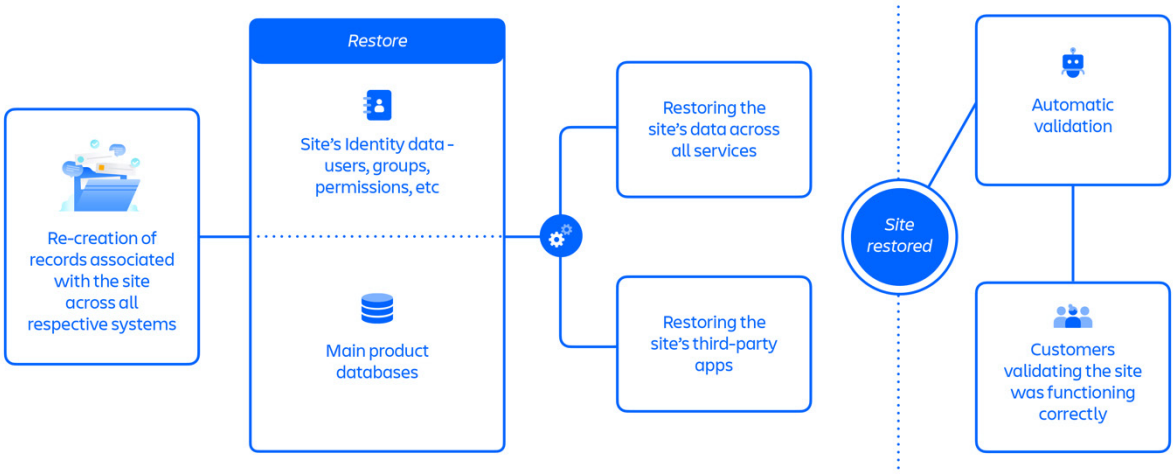
如果使用「復原方法 1」，會需要三週的時間才能為所有客戶復原網站。因此我們在 4 月 9 日提出一項新方法，以加快復原所有網站，這項方法為「復原方法 2」(請參閱 *下方圖表 8*)。

「復原方法 2」藉由降低複雜度以及降低「復原方法 1」的相依性數量，改善了復原步驟之間的並行性。

「復原方法 2」需要從目錄服務記錄開始，在所有個別系統中重新建立 (或取消刪除) 與網站相關的記錄。這項新方法的一個關鍵要素是*重複使用所有舊的網站識別碼*。之前的方法需要對應新舊識別碼。「復原方法 2」免除了至少一半的對應步驟，而且無須協調每個網站的第三方應用程式供應商。

不過，要將「復原方法 1」轉換為「復原方法 2」，為事件回應增加了大量額外負荷：

- 在「復原方法 1」中建立的許多自動化指令碼與流程都必須針對「復原方法 2」進行修改。
- 我們在測試並驗證「復原方法 2」的流程時，執行復原的團隊 (包含事件協調員) 必須運用這兩種方法，管理並行的批次復原。
- 要使用新復原方法，代表我們需要在擴展工作前測試並驗證「復原方法 2」的流程，也就是需要複製之前為「復原方法 1」完成的驗證工作。



圖表 8：復原方法 2 的重要步驟。

上方圖示代表了「復原方法 2」，包含 30 多個步驟，這些步驟遵循以下並行流程：

- 在所有個別系統中重新建立與網站相關的記錄
- 復原網站的身份識別資料- 使用者、群組、權限等
- 復原網站的主要產品資料庫
- 在所有服務中復原網站的資料
- 復原網站的第三方應用程式
- 自動驗證
- 客戶驗證網站是否能正常運作

為了加速復原，我們也採取了前期吃重與網站復原自動化等步驟，因為手動復原無法隨著大批次的工作進行擴展。復原流程的順序表示大型資料庫復原、使用者群/權限復原的網站復原工作可能會比較緩慢。我們實施的最佳化包括：

- 我們開發了*前期吃重*所需的工具和規範，以及長期執行步驟 (例如資料庫修復和身分識別同步)，可在其他修復步驟之前完成。
- 工程團隊為各自的步驟構建了自動化，從而使大批次的修復能夠安全地執行。
- 構建自動化是為了在所有修復步驟完成後，驗證網站是否能正常運行。

加速的第 2 種修復方法大約需要 12 個小時來修復網站。在 4 月 14 日到 17 日這段期間，在 771 個網站上受影響的使用者中，大約有 47% 是利用這種方法進行修復。

修復遭到刪除的網站後，所遺失的資料最少

我們的資料庫備份方式結合了完整備份和增量備份，讓我們能夠在備份保留期間 (30 天) 內擇選任何特定的「時間點」，復原資料儲存區。針對大多數客戶，在此事件期間我們找出了自家產品的主要資料儲存區，並決定使用網站遭到刪除前的 5 分鐘這個還原點，來作為安全同步點。非主要資料儲存區會還原到同一個還原點，或在重播記錄下來的事件後，決定要還原到哪個還原點。對主要儲存區使用固定還原點使我們能夠跨所有資料儲存區，取得資料的一致性。

對於我們在事件的回應初期還原的 57 位客戶，由於缺乏一致的政策，且需要手動取得資料庫備份快照，導致一些 Confluence 和 Insight 資料庫，被還原到比網站遭到刪除前 5 分鐘*更早*的還原點。在修復後的審核過程中，我們才發現這種不一致的地方。接著，我們復原了剩餘的資料、連絡了受此影響的客戶，到目前為止仍然持續協助他們套用變更，以進一步還原資料。

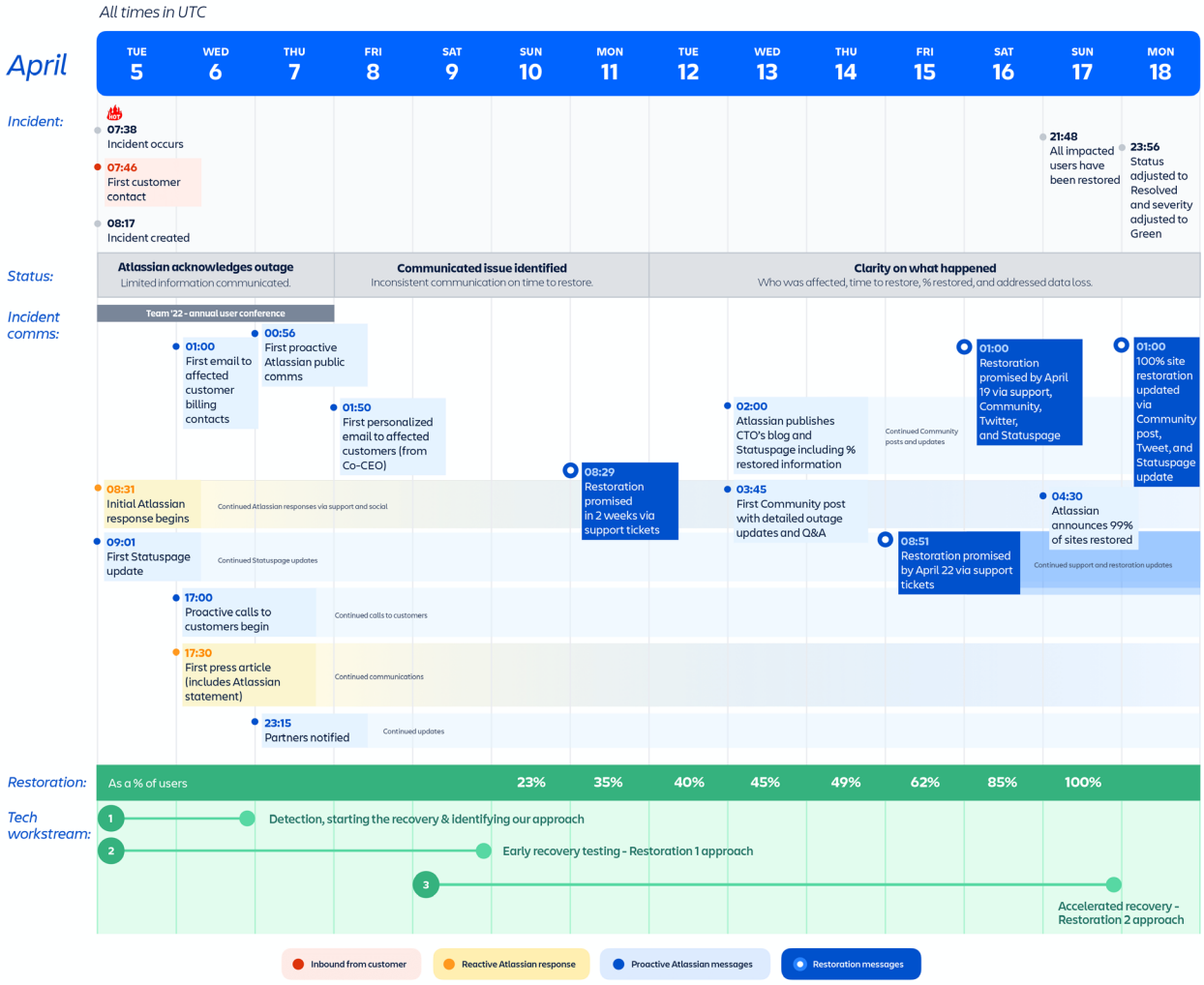
總而言之：

- 在此事件期間，我們達到了一個小時的復原點目標 (RPO)。
- 在事件中所遺失的資料範圍，已限制在網站遭到刪除前的 5 分鐘。
- 有少數客戶的 Confluence 或 Insight 資料庫，被還原到比網站遭到刪除前 5 分鐘更早的還原點，不過，我們能夠復原這些資料，目前也正在與客戶一起努力還原這些資料。

事件溝通

事件溝通包括了與客戶、合作夥伴、媒體、業界分析師、投資者和廣大的技術社群等接觸點。

事件始末



圖表 9：重要事件溝通里程碑的時間表。

時間戳記：第 1 天到第 3 天 (4 月 5 日到 7 日)

初期回應

第一份支援工單是在 4 月 5 日 7:46 時建立的，而 Atlassian 支援團隊在 8:31 前回應已承認此事件的發生。到 9:03 時，我們發佈了第一個 Statuspage 更新，讓客戶知道我們正在調查此事件。到 11:13 時，我們透過 Statuspage 確認已找出根本原因，並設法修復問題。到 4 月 6 日 1:00 時，最初的客戶支援工單通訊顯示停機事件是因為維護指令碼而發生，且我們預期遺失的資料會在最小限度。在 4 月 6 日 17:30 時，Atlassian 發表了聲明，以回覆媒體的詢問。在 4 月 7 日 00:56 時，Atlassian 發佈了第一份廣泛的外部訊息推文，承認此事件的發生。

時間戳記：第 4 天到第 7 天 (4 月 8 日到 11 日)

開始 更廣泛、更專屬的對外連絡

在 4 月 8 日 1:50 時，Atlassian 的共同創辦人暨共同執行長 Scott Farquhar 透過電子郵件向受影響的客戶致歉。在接下來的幾天裡，我們設法還原了遭到刪除的連絡資訊，也為所有受影響但尚未提交支援工單的網站建立了支援工單。接著，我們的支援團隊繼續透過與各個受影響的網站相關的支援工單，為修復網站提供了定期更新。

時間戳記：第 8 天到第 14 天 (4 月 12 日到 18 日)

事情更加明朗，修復也終於完成

在 4 月 12 日， Atlassian 的技術長 Sri Viswanath 發佈了更新，提供這次事件更多的技術細節、受影響的範圍、是否遺失了資料、修復的進度，以及說明最多會需要兩週的時間才能還原所有網站。除了這個部落格文章，Sri 還發表了其他媒體聲明。我們的工程主管 Stephen Deasy 第一次主動在 Atlassian 社群貼文時，也提到了 Sri 的部落格，之後這個部落格就成了發表其他更新和大眾進行問答的專用場所。4 月 18 日這個貼文的更新中，宣告了所有受影響客戶的網站都已完全修復。

為什麼我們沒有更早公開回應？

1. 我們認為透過 Statuspage、電子郵件、支援工單和一對一與受影響的客戶直接溝通，是最重要的。但是，我們仍無法聯繫上許多客戶，因為在他們的網站遭到刪除時，我們也失去了他們的連絡資訊。我們應該更早進行更廣泛的溝通，以便告知受影響客戶和終端使用者我們對事件的回應和解決時間表。

2. 雖然我們立即知道事件發生的原因是什麼，但這次事件在架構上的複雜性和特殊情況，在在讓我們難以快速確定範圍和準確估計解決時間。我們應該向客戶坦承我們知道什麼、又不知道什麼，而不是等到了解全貌後才告知客戶。提供一般的修復進度估計 (即使只是大概的方向)，並清楚傳達我們預計何時能對事件情況有更透徹的了解，都能讓我們的客戶能據此擬定計畫。對於系統管理員和技術團隊連絡人來說尤其如此，因為他們處於管理組織內相關人員和使用者的最前線。

支援體驗與客戶推廣

如前所述，刪除了客戶網站的指令碼，也從我們的正式環境中，刪除了關鍵的客戶識別碼和連絡資訊 (例如：Cloud URL、網站系統管理員連絡人)。我們能發現這點，是因為我們的核心系統 (例如：支援、授權、計費) 全都利用 Cloud URL 和網站系統管理員連絡人，作為安全、路由和優先順序的主要識別碼。所以當我們失去了這些識別碼時，我們一開始失去了系統化地識別客戶並與客戶互動的能力。

我們對受影響的客戶提供了哪些支援？

首先，大多數受影響的客戶無法透過正常的線上連絡表單聯繫我們的支援團隊。此表單旨在要求使用者以自己的 Atlassian ID 登入並提供有效的 Cloud URL。使用者如果沒有有效的 URL，就無法提交技術支援工單。在正常業務過程中，是為了顧及網站安全和支援工單分級而特意進行此驗證。但是，這項要求對受到停機事件影響的客戶造成了意外的結果，也就是無法提交高優先順序的網站支援工單。

其次，事件導致網站系統管理員的資料遭到刪除，讓我們難以主動與受影響的客戶聯繫。在事件發生的頭幾天，我們主動聯繫了在 Atlassian 註冊的受影響客戶，與他們的帳務連絡人和技術團隊連絡人聯繫。但我們很快發現，受影響客戶的許多帳務連絡人和技術團隊連絡人資料早已過時。如果沒有每個網站的系統管理員資訊，我們就無法取得經過核准的活躍連絡人完整列表，因此無法進行聯繫。

我們如何回應？

在事件發生的頭幾天，為了加速修復網站，及修復通訊管道中遭到破壞的地方，我們的支援團隊有三個同樣重要的優先要務。

首先，要取得經過驗證而可靠的客戶連絡人清單。在我們的工程團隊努力還原客戶網站時，我們要面對客戶的各團隊也把重點放在還原經過驗證的連絡資訊上。我們使用了手頭上所有可用的方法 (帳務系統、先前的支援工單、其他安全的使用者備份、直接對外連絡客戶等)，來重建連絡人清單。我們的目標是為每個受影響的網站建立一個與事件相關的支援工單，以便縮短直接對外連絡和回應的時間。

其次，要重新建立此事件特定的工作流程、佇列和 SLA。Cloud ID 遭到刪除，又無法正確驗證使用者，也影響了我們透過正常系統處理事件相關支援工單的能力。支援工單所顯示的相關優先順序、呈報佇列和儀表板全都不正確。我們快速籌組了一個跨職務團隊 (來自支援、產品、IT 等領域)，設計並新增了額外的邏輯、SLA、工作流程狀態和儀表板。因為這必須在我們的正式作業環境系統中完成，所以需要幾天時間才能完成全面開發、測試和部署。

第三，大規模擴充手動驗證的範圍，以加速修復網站。隨著工程團隊在一開始的還原過程中取得進展，顯然會需要我們的全球支援團隊奧援，透過手動測試和驗證檢查來加速修復網站。一旦我們的工程團隊還原資料的速度加快，此驗證流程將成為還原客戶網站的關鍵途徑。我們必須建立獨立的標準操作程序 (SOP)、工作流程、移交和人員輪值表等流程，以動員超過 450 名的支援工程師進行驗證檢查，全天候輪班不打烊，讓資料能夠更快回到客戶的手中。

即使在第一週結束時已經確立了這些關鍵優先事項，但由於修復流程的複雜性，事件解決時間表仍不夠明朗，我們能提供的*有意義*更新也因此受限。我們應該更早坦承自己不確定何時能修復網站，並準備好進行面對面討論，讓我們的客戶能據此擬定計畫。

我們將如何改進？

我們已立即禁止大量網站刪除，直到可以進行適當的變更為止。

在我們擺脫這次事件的影響，並重新評估內部流程後，我們認定事件不是由人造成的，而是因為系統的疏漏，才會導致錯誤發生。本節總結了造成此事件的因素。還探討了我們要如何加速解決這些弱點和問題的計畫。

心得 1：「軟刪除」應該在所有系統中普遍使用

整體而言，本事件類型的刪除方式應受到禁止，或者具有多層保護以避免發生錯誤。我們正在進行的主要改進，是全面避免尚未經過軟刪除流程的客戶資料和中繼資料遭到刪除。

a) 只能使用軟刪除方式進行刪除

應禁止刪除整個網站；即使是軟刪除，也需要多層保護以避免錯誤。我們將實施「軟刪除」政策，防止外部指令碼或系統刪除客戶在正式作業環境中的資料。我們的「軟刪除」政策將允許系統保留足夠的資料，以便快速安全地執行資料復原。只有在保留期過後，資料才會從正式作業環境中刪除。

行動：

- **在佈建工作流程及所有相關的資料儲存區中，實施「軟刪除」：**此外，租戶平台團隊會確認只有在停用後才能刪除資料，在此空間也會設有其他的保障措施。長期來說，租戶平台在進一步開發正確的租戶資料狀態管理機制時，會發揮主導作用。

b) 軟刪除應具有標準化且經過驗證的審查流程

軟刪除動作是高風險的操作。因此，我們應設有標準化或自動化的審核流程，包括定義好的復原和測試程序，來因應這些操作。

行動：

- **強制分階段推型任何軟刪除動作：**所有需要刪除的新操作，首先將在我們自己的網站中進行測試，以驗證我們的方法是否可行和自動化是否有效。完成驗證後，我們將逐步協助客戶推動相同的流程，並繼續測試不適當的地方，然後再將自動化流程套用到整個所選的使用者群體。

- **軟刪除動作必須具有經過測試的復原計畫：**任何會軟刪除資料的活動，在正式作業環境系統中執行前，都必須先測試能否修復刪除的資料，且必須具有經過測試的復原計畫。

心得 2：作為災難復原計畫的一部分，在大群客戶遭遇多網站、多產品的刪除事件時，應能進行自動化修復

Atlassian 資料管理詳細介紹了我們的資料管理流程。為確保高度可用性，我們在多個 AWS Availability Zones (AZ) 中佈建並維護同步備用複本。AZ 的容錯移轉是自動的，通常需要 60-120 秒的時間。我們會定期處理資料中心停機和其他常見中斷，這不會對客戶造成影響。

我們同時也維護不可變備份，這些備份是可復原的，若遇到資料損壞事件，能夠復原到上一次的正確狀態。備份會保留 30 天，Atlassian 會持續測試和審核備份儲存，以供修復之用。若有需要，我們也可以一次將所有客戶恢復到新的環境中。

有了這些備份，我們就可以定期為誤刪資料的個人客戶或少數客戶復原資料。但是，網站級別的刪除沒有能夠快速自動執行此事件規模的運行手冊，這需要以協調的方式跨所有產品和服務進行工具和自動化。

我們尚未自動化的部分是在不影響其他客戶的情況下，將一大部分客戶恢復到我們現有使用的環境中。

在我們的雲端環境中，每個資料儲存區都包含來自多位客戶的資料。由於此事件中遭到刪除的資料只是其他客戶持續使用的一部分資料儲存區，所以我們必須手動從備份中一一擷取並還原資料。每個客戶網站的復原過程皆耗時又複雜，需要在修復網站時進行內部驗證和最終客戶驗證。

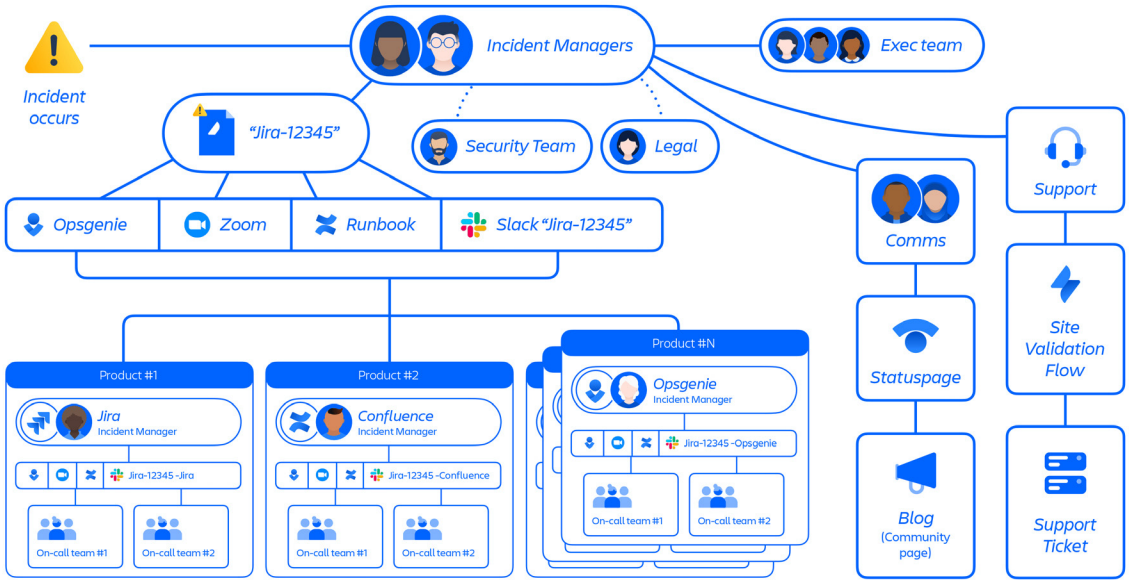
行動：

- **為更多的客戶加快多網站、多產品的修復速度：**災難復原計畫需滿足我們當前的一小時 RPO 標準。我們將利用這次事件中學到的教訓和自動化經驗，加快災難復原計畫的執行速度，以滿足我們的策略中，針對此規模事件所定義的 RTO。
- **推行自動化，將此案例中的驗證方式新增到災難復原的測試中：**我們會定期執行災難復原演練，練習回復大量網站的所有產品。這些災難復原測試，將驗證我們的標準執行程序是否能隨著我們的架構發展與時俱進，以及是否會碰到極端的案例。我們將不斷改進修復方法，將修復流程更加自動化，並縮短復原時間。

心得 3：改進大規模事件的事件管理流程

我們的事件管理計畫，非常適合管理經年累月下來發生的大大小小事件。我們經常對較小規模、持續時間較短的事件進行模擬事件回應，這些事件通常涉及較少的人員和團隊。

但是，在此事件的高峰期，有數百名工程師和客戶支援員工同時工作，以恢復客戶網站。事件管理計畫和團隊的設立目的並不是要處理如此深入、廣泛、影響持久的事件 (請參閱 *下方圖表 10*)。



圖表 10：大規模事件管理流程的概覽。

我們會定義出更好的大規模事件管理流程並勤加演練

我們有產品層級事件的教戰守則，但沒有這種規模的事件使用案例，也就是公司內的數百位員工同時工作的規模。在我們的事件管理工具中，有自動化功能可以建立像 Slack、Zoom 和 Confluence 文件這樣的溝通管道，但缺乏大規模事件所需的分支管道，因此無法將修復所需的溝通管道獨立出來。

行動：

- **為大規模事件定義教戰守則和工具，並進行模擬練習：**定義並記錄可能被視為大規模且需要這種等級回應的事件類型。概述關鍵協調步驟並構建工具，以協助事件管理員和其他職務的人員簡化回應流程並開始復原。事件管理員及其團隊將定期進行模擬和訓練，並改善工具和文件，以持續改進。

心得 4：改進我們的溝通流程

a) 我們不慎刪除了重要的客戶識別碼，與受影響客戶之間的溝通和我們能夠採取的行動也因此受到影響

刪除了客戶網站的指令碼，也從我們的正式作業環境中，刪除了關鍵的客戶識別碼 (例如：網站 URL、網站的系統管理員連絡人)。因此，(1) 客戶無法透過正常的支援管道提交技術支援工單；(2) 因為喪失了主動接觸客戶的能力，我們花了好幾天時間才取得可靠的關鍵客戶連絡人 (例如網站的系統管理員) 列表；以及 (3) 因為這次事件的獨特性質，一開始支援的工作流程、SLA、儀表板和呈報流程都無法運作。

在停機期間，客戶也透過多種管道進行呈報 (電子郵件、電話、執行長工單、LinkedIn 和其他社交媒體管道和支援工單)。要面對客戶的各團隊，各自接到了來自不同工具和流程的呈報，不但拖慢了我們回應的速度，也讓整體的追蹤和報告變得更加困難。

b) 我們當時手頭上的事件溝通教戰守則，都沒有詳盡到足以處理複雜到這種程度的事件

我們缺乏概述了原則、職位和責任的事件溝通教戰守則，能夠及時跨職務統一動員整個團隊。我們未透過多種管道，尤其是社交媒體，迅速並一致地承認此事件的發生。發生停機事件後，透過更加廣泛、公開的方式與客戶溝通，並重複告知此事件並未造成任何資料遺失，而且也不是網路攻擊所造成，才是正確的做法。

行動：

- **改進關鍵連絡人的備份：**備份產品執行個體以外的授權帳戶連絡資訊。

- **改良支援工具：**為沒有有效 URL 或 Atlassian ID 的客戶，建立直接與我們技術支援團隊連絡的機制。

- **客戶呈報系統和流程：**投資在以帳戶為基礎的統一呈報系統和工作流程上，可將多個工作物件 (支援工單、任務等) 儲存在單一客戶的帳戶物件下，讓我們要面對客戶的各團隊更能夠進行協調，也更能夠對這些物件一目了然。

- **加快全天候呈報管理的涵蓋範圍：**在全球使用量擴展計畫中執行呈報管理功能，可讓全球得到一致的全天候服務，每個主要地理區域都有專屬的員工以及支援職位，為客戶所需的產品、銷售題材的專家和領導人才等提供協助。

- **利用新學到的經驗更新我們的事件溝通教戰守則，並定期溫故知新：**溫習教戰守則，在內部定義明確的職位和界線。使用 DACI 架構處理事件，並為每個職位提供全天候的職務代理機制，以因應員工因為生病、度假或其他原因請假的情況。每季進行審核，隨時驗證是否準備好應付一切狀況。

行動 (續)

在所有溝通中遵循事件溝通範本：提到發生的事件、受影響的範圍、

修復的時間表、網站修復的百分比、預期會遺失的資料多寡、相關的信心水準，

以及如何連絡支持團隊的明確指示。

結語

雖然停機事件已解決，客戶的資料也已經完全還原，但我們的工作並未因此停止。在此階段，我們正在實施上述的各項變革，以改進我們的流程、增強我們對突發事件的抵抗力，並防止這種情況再次發生。

Atlassian 是一個不斷學習、不斷進化的組織，我們的團隊肯定從這次的經驗中學到了不少慘痛的教訓。我們會記取這些教訓並加以活用，讓我們的業務能發生持久的變革。最終，我們會因為這次的經驗而更加強大，並能為您提供更好的服務。

我們希望從這次事件中獲得的經驗，對致力於為客戶提供可靠服務的其他團隊也有所幫助。

最後，我要感謝那些正在閱讀這篇文章並與我們一起學習的人，以及我們 Atlassian 的社群和團隊大家族。

- 技術長 Sri Viswanath

ATLASSIAN