前言

近年來，遷移到雲端的便捷性大為改善，並使得快速增長的資料消費者社群，在蒐集、捕獲、存儲和分析資料以獲得洞察力和決策制定方面，變得充滿更多可能性。由於各種原因，並隨著雲端計算的採用不斷地增長，資訊管理利益相關者，開始對在雲中管理資料所涉及的潛在風險有所疑問。從事醫療保健行業工作的 Evren 第一次遇到這樣的問題，而不得不制定流程和技術以管理資料。現今，在 Google Cloud（Google 雲 / Google 雲端）上，Uri 和 Lak 幾乎每週都會回答這些問題，並就從資料中獲取價值、打破資料孤島、保持匿名、保護敏感資訊和提高資料可信度等方面提出建議。

我們注意到 GDPR[1] 是讓客戶行為徹底變化的原因，其中我們的一些客戶甚至認為，他們必須刪除所擁有的資料，以符合 GDPR 規定。和其他反應相比起來，正是這種反應更促使我們要寫這本書，以藉此記錄我們多年來向 Google Cloud 客戶提供的建議。如果資料是新興貨幣，我們不希望企業對它感到恐懼；資料一旦遭到鎖定或不受信任，就沒有價值。

我們都為幫助 Google Cloud 客戶獲得技術支出的價值而感到自豪。資料是一項巨大的投資，我們自覺有義務為客戶提供從中獲取價值的最佳方式。

1 *https://gdpr.eu/what-is-gdpr*

客戶的問題通常不外乎以下三個風險因素：

保護資料

在本地端部署系統的大型企業通常期望嚴謹的安全性，因此，將資料儲存在公有雲的基礎設施中，往往會使這些企業對安全性有所疑慮。隨著新聞中出現大量的安全威脅和資料洩露事件，也讓組織擔心他們可能會成為下一個受害者。這些因素帶來了風險管理問題，如防止未經授權的存取或暴露敏感資料，這些敏感資料包括個人識別資訊（PII）以及公司機密資訊、商業機密或知識產權。

法規和合規

近年來，一系列不斷增長的法規日益引起關注，包括加州消費者隱私保護法案（CCPA）、歐盟的一般資料保護規定（GDPR）以及某些行業的特定標準，例如金融行業的全球法人機構識別碼（LEI）和保險行業的資料標準ACORD。而負責遵守這些法規和標準的合規團隊，可能會對監督和控制存儲在雲中的資料感到憂心。

可視性與其控制

資料管理專業人員和資料消費者有時無法了解自己身處的資料環境：哪些資料資產可用？這些資產位於何處以及如何使用和是否可以使用？誰有權存取資料以及他們是否應該擁有對資料的權限？這種不確定性限制了他們進一步地利用自己的資料，來提高生產力或推動業務價值的能力。

這些風險因素清楚地表明，需要增加資料評估、替元資料編寫目錄、存取控制管理、資料品質和資訊安全以作為核心資料治理能力。雲端供應商不僅應該提供這些能力，而且還應該以透明的方式不斷地為其升級。從本質上講，在不放棄雲端計算提供好處的情況下解決這些風險，不僅更能讓人理解在雲端中資料治理的重要性，而且也能更加地理解哪些資料較為重要。良好的資料治理可以讓客戶感到信任，並有效改善客戶體驗。

為什麼您的企業需要雲端中的資料治理？

隨著您的企業產生更多資料並將其移動到雲中，資料的管理會在許多基本方面發生變化。組織應注意以下事項：

風險管理

敏感資訊暴露給未經授權的個人或系統、安全漏洞，或已知人員在錯誤的情況下存取資料等，都讓人擔心。有鑑於此，組織希望將這種風險降至最低，因此需要額外的保護形式，例如加密，以混淆資料物件內的嵌入式資訊，使得在系統遭到破壞時也能夠保護資料。此外，還需要其他工具來支持存取管理、識別敏感資料資產並制定保護策略。

資料的激增

企業建立、更新和串流傳輸其資料資產的速度有所提高，雖然基於雲端的平台能夠處理不斷增加的資料、數量和種類，但在高頻寬串流資料方面，引入控制和相關機制以快速地驗證品質非常重要。

資料的管理

若是需要採用外部的資料來源和串流資料，包括來自第三方的付費資料，即意味著您應該有心理準備，不信任所有外部資料來源。您可能需要引入記錄資料歷程、分類和元資料的工具，以幫助您的員工（尤其是資料消費者）瞭解所使用的資料，並且根據他們對資料資產生成方式的了解，來確定資料的可用性。

資料的發現（及資料感知）

將資料移動到任何類型的資料湖泊，不管是雲端或本地端，都存在失去追蹤哪些資料資產已移動、其內容的特徵以及元資料細節的風險。因此，無論該資料位於何處，具備評估資料資產內容和其敏感性的能力變得非常重要。

隱私性和合規性

法規遵從性需要可審計和可衡量的標準與程序，以確保遵守組織內部的資料政策以及政府的法規。並且，將資料遷移到雲端上意味著組織需要工具來執行、監控和回報合規性，並確保正確的資料能交給對的人，使用在對的事情上。

雲端中的資料治理框架和最佳實踐

有鑑於資料管理經常處於變化中，組織應如何思考在雲端中的資料治理及其重要性？根據 *TechTarget* 的說法，資料治理是：

> 對企業使用的資料而言，是指可用性、易用性、完整性和安全性的整體管理。一個健全的資料治理計畫，包括一個管理機構或委員會、一套明確的程序和執行這些程序的計畫[2]。

簡而言之，資料治理包含人員、流程和可以協同工作的技術、方式，針對已定義和具備共識的資料治理策略，實現可被審計的合規性。

資料治理框架

企業需要通盤性的考量資料治理，從攝取何種資料、攝入多少資料量到編撰目錄、持久化、保留、存儲管理、共享、歸檔、備份、恢復、預防資料遺失、配置以及移除和刪除：

資料的發現和評估

當談到建立和管理資料湖泊時，使用雲端環境通常代表提供的是一種經濟導向的選擇，但針對資料資產的遷移，仍存在其不受監管的風險。這種風險代表一種潛在損失，指的是對資料湖泊中資料資產的了解程度、每個物件中包含的資訊以及這些資料物件的來源。雲端中資料治理的最佳實踐是資料的發現和評估，以便了解您擁有哪些資料資產。資料發現和評估流程用於識別雲端環境中的資料資產，並且追蹤和記錄每個資料資產的來源和歷程、應用了哪些轉換以及該物件元資料。（這裡對元資料描述的資訊類似人口統計學科那樣詳細，例如創建者姓名、物件大小、如果是結構化資料物件的話，則記錄筆數；或該物件的最近更新時間。）

2 Craig Stedman 和 Jack Vaughan，〈何謂資料治理及其重要性？〉（What Is Data Governance and Why Does It Matter?）TechTarget，2019 年 12 月。本文於 2020 年 2 月更新；當前版本已不再包含此引用（*https://oreil.ly/OdvVk*）。

資料分類和組織

正確地評估資料資產，並掃描不同屬性的內容有助於對資料資產進行後續的組織分類工作。此過程還可以推斷所分析之物件是否包含敏感資料，如果包含敏感資料，如個人及私人資料、機密資料或知識產權，則需根據該資料敏感級別分類。若要在雲中實施資料治理，您需要概要分析和分類敏感資料，以確定哪些治理策略和程序適用於該資料。

編撰資料目錄及元資料管理

一旦您的資料資產可評估和分類，則記錄您的學習收穫至關重要，這樣您的資料消費者社群就可以了解您所組織的資料樣貌。有鑑於此，您需要維護一個資料目錄，其中包含結構化的元資料、物件元資料，以及與治理指令相關的敏感度級別評估（例如遵守一項或多項資料隱私法規）。資料目錄不僅允許資料消費者查看此資訊，而且還可以作為搜索和發現功能的反向索引的一部分，只要給定正確關鍵字不管是按特定片語或是某個概念名詞均可以搜索。此外，了解資料物件是屬於結構化或半結構化格式也很重要，因為您需要依據實際需求，以允許系統分別地處理不同格式的資料物件。

資料品質管理

當談論資料品質時，不同的資料消費者可能會有不同的要求，因此提供一種符合期望的記錄資料品質方法，以及支持資料驗證和監控過程的技術和工具非常重要。資料品質管理流程包括建立驗證控制、啟用品質監控和報告、支持用於評估事件嚴重程度的分類流程、啟用根本原因分析和資料問題補救措施建議，以及資料事件追蹤。正確的資料品質管理流程將提供可衡量且值得信賴的資料，以便分析。

資料存取管理

關於資料存取的治理有兩個方面。第一個方面是提供對可用資產的訪問。提供資料服務以允許資料消費者存取十分重要，幸運的是，大多數雲服務平台都提供開發資料服務的方法。第二個方面是防止不當或未經授權的存取。定義身分、分組和角色並分配存取權限，以建立一定程度的存取託管也很重要。這樣的最佳實踐涉及管理存取服務，以及透過定義角色、指定存取權限、管理和分配存取金鑰，來與雲端服務供應商的身分識別與存取管理（IAM）服務進行互動操作，這些措施得以確保只有經過授權和身分驗證的個人和系統，才能根據已定義的規則存取資料資產。

審計

組織必須要能夠評估他們的系統，以確保系統按照設計初衷運行。因此，監視、審計和追蹤，即知道誰在何時做什麼事，以及使用哪些資訊，能幫助安全團隊蒐集資料、識別威脅，並在威脅導致業務損害或發生損失之前採取行動。定期執行審計以檢查控制措施的有效性非常重要，以便快速應對威脅並評估整體安全狀況。

資料保護

儘管資訊技術安全小組努力建立外圍安全措施，以作為防止未經授權的個人存取資料的一種方式，但外圍安全措施一直以來都不足以保護敏感資料。雖然您可能成功地阻止某人闖入您的系統，但您仍然無法防止內部安全漏洞甚至資料洩露。因此，制定額外的資料保護，包括靜態加密、傳輸中加密、資料屏蔽和永久刪除等方法非常重要，以確保已遭受暴露的資料無法讀取。

在您的組織中實施資料治理

相關技術當然有助於支持上一節介紹的資料治理原則，但是實務上，資料治理遠超出了產品和工具的選擇與實施範圍。資料治理計畫的成功取決於以下因素的組合：

- 建立業務案例、開發營運模型並承擔適當角色的**人員**
- 可行的政策、實施和執行的**流程**
- 幫助人們執行這些流程方式的**技術**

以下步驟對於規劃、啟動和支持資料治理計畫至關重要：

1. **構建業務案例**。藉由識別關鍵業務驅動因素來建立業務案例，以證明與資料治理相關的工作和投資是合理的。概述那些您感知到的資料風險（例如在雲端平台上存儲資料），並說明資料治理如何幫助組織減輕這些風險。
2. **文件指導原則**。確立與企業資料治理和監督相關的核心原則。在資料治理章程中記錄這些原則，以提交給管理層。
3. **獲得管理層的支持**。吸引資料治理擁護者，並獲得主要高級利益相關者的支持。將您的業務案例和指導原則提交給公司高層以供批准。

4. **開發營運模式**。獲得管理層批准後，定義資料治理角色和職責，然後為資料治理委員會和資料管理團隊描述流程和程序，他們將制定相關流程，以用於定義和實施策略，以及審查和補救已識別資料的問題。

5. **建立當責框架**。建立一個框架來分配關鍵資料領域的保管和責任。確保資料環境中的「資料擁有者」對其資料的可見性。並提供一種方法，以確保每個人都對資料可用性的貢獻負責。

6. **發展分類方式和本體**。考量在治理實踐上，可能有許多與資料分類、組織，以及在敏感資訊情況下的資料保護相關治理指令。為了使您的資料消費者能夠遵守這些指令，必須明確定義類別（用於組織結構），和分類（用於評估資料敏感性）。

7. **集結正確的技術堆棧**。一旦您為員工分配資料治理角色，並定義和批准您的流程和程序，就應該組裝一套工具，來促進持續驗證資料策略的合規性和準確的合規性報告。

8. **建立教育系統和培訓機制**。藉由發展教育用途的相關材料，以強調資料治理的實踐、程序，和支持治理的技術使用，提高對資料治理價值的認識。照計畫定期培訓課程，以加強良好的資料治理實踐。

強大資料治理的商業利益

資料安全、資料保護、資料可存取性，和易用性、資料品質以及資料治理的其他方面將陸續誕生，並發展成為組織的關鍵優先事項。隨著越來越多的組織將其資料資產遷移到雲端，對確保資料實用性的可審計實踐需求也將繼續增長。為了解決這些面向，企業應圍繞著三個關鍵組成部分，以制定資料治理實踐：

- 一個使人們能夠定義、同意和執行資料策略的框架
- 跨本地系統、雲端存儲和資料倉儲平台，控制、監督和管理所有資料資產的有效流程
- 用於實施資料策略合規性的正確工具和技術

考量到這個框架，有效的資料治理政策和營運模型為組織提供了建立控制和保持資料資產可見性的途徑，從而提供了優於同行的競爭優勢。當組織在其內部推廣資料驅動的文化時，它們可能會獲取許多好處，尤其是：

改進決策制定

更完善的資料發現，意味著使用者永遠可以在需要時找到資料，從而提高效率。資料驅動的決策制定在改進組織內的業務規劃方面發揮重要作用。

更理想的風險管理

良好的資料治理運營模式可幫助組織更輕鬆地審核其流程，從而降低被罰款風險、增加客戶信任並改善營運。可以最大限度地減少停機時間，同時提高生產率。

監管合規性

越來越多的政府監管使得組織建立資料治理實踐變得更加重要。有了良好的資料治理框架，組織可以適應不斷變化的監管環境，而不只是簡單地對其做出反應。

隨著您將更多資料遷移到雲端中，資料治理提供了一定程度的資料濫用保護。同時，對已定義的資料策略的可審計合規性，有助於向您的客戶證明您能保護他們的私人資訊，減輕他們對這方面風險的擔憂。

本書目標讀者

當前的資料成長幅度前所未有，加上法規和罰款的增加，意味著組織被迫研究自身資料治理計畫，以確保它們不會成為下一個被處罰的對象。因此，每個組織都需要了解其蒐集的資料內容、與該資料相關的責任和法規，以及有權存取這些資料的人。如果您想知道這代表的意義、需要注意的風險以及必須牢記的注意事項，則這本書適合您閱讀。

本書也適用於那些需要讓資料變得更為可信的流程或技術的人。本書涵蓋人員、流程和技術可以協同工作的方式，針對已定義和具備共識的資料治理策略，實現可被審計的合規性。

資料治理的好處為多面向，從法規、合規性到更好的風險管理，以及藉由建立新產品和服務，來推動營收和節約成本的能力。閱讀本書，了解如何建立對資料資產的控制並保持可見性，將能為您提供超越同行的競爭優勢。

第一章

何謂資料治理？

資料治理是一種資料管理功能，用於確保組織所蒐集的資料質量、完整性、安全性和可用性。從資料的生成或蒐集，直至該資料的銷毀或備份存檔，每個時間點都需要考慮置入資料治理。在資料生命週期中，資料治理的重點是提供某種形式，使所有的利益相關者都可以輕鬆地存取資料。此外，這個工具必須是他們可藉此產生所需業務成果（某種洞見、分析），並符合相關監管標準的方式。而這標準通常代表某些實體單位的共識與交集，如醫療保健業；或政府所需的資料隱私性，及一般公司行號的政治立場中立等。此外，資料治理需要確保企業內所有資料能整合成高品質的資料，讓所有利益相關者獲取參考。高品質資料的定義有很多，例如正確、最新且一致的資料。最後，置入資料治理是為了確保資料安全，也就是說：

- 只允許得到授權的使用者，以合法方式存取資料
- 它是可受稽核的，這意味著所有存取，包括更改都會記錄下來
- 確保一切符合規範

資料治理的目的是增強對資料的信任。值得信賴的企業資料必不可少，它讓使用者能夠藉此來支持決策的制定、風險的評估和使用關鍵績效指標（KPI）以管理。藉由資料的使用，您手上有足夠的證據可以加強對決策過程的信心。並且，無論企業規模大小、資料量多寡，資料治理的原則都是一樣的。然而，參與者往往會考慮到他們所處的實際環境，在資料治理的工具層面和實作方面做出選擇。

資料治理的內容為何？

巨量資料分析的出現，得益於將資料遷移到公有雲的便捷性，和不斷增長的計算能力，並進一步驅動一群雄心壯志的資料工作者蒐集、存儲和分析資料，以獲取洞見與決策。如今，幾乎每個電腦應用程式，都是基於蒐集而來的商業資料以提供資訊服務，因此，無論是藉由新系統以蒐集資料集，還是從外部供應商購買資料集，新點子的產生，不可避免地涉及以新的方式分析既有資料。您的組織是否有機制來審查新的資料分析技術、確保蒐集到的所有資料都得到安全存儲且具備一定品質？而且，基於此資料而產生的功能，是否會增加您的品牌價值？雖然一般人很容易只關注資料蒐集和巨量資料分析的威力和可能性，但資料治理是一個非常現實且再重要不過的考慮因素，不容忽視。2017 年，《哈佛商業評論》（*Harvard Business Review*）指出，超過 70% 的員工可以存取他們不應該存取的資料[1]。這並不是說公司應該採取防禦姿態；這只是說明了，資料治理對於防止資料洩露和不當使用的重要性。良好的資料管理可以為組織帶來可衡量的收益。

Spotify 創造新功能：個人化音樂播放清單

說到良好的資料管理能為組織帶來的可衡量收益，以及徹底改變整個行業的資料可用性，最好的例子就是 Spotify 的個人化音樂清單功能。在 2010 年代初期，大多數人聽音樂的方式仍然是購買實體專輯，或是將擁有的歌曲、最近收聽的內容截錄至電腦，以重新編排，創建屬於自己的播放清單。

除了合法管道取得歌曲以外，還有一個龐大且繁榮的非法音樂共享生態系統，讓人可以添加盜版音樂至播放清單中。而為了抑制盜版音樂生態，唱片公司便允許音樂以數位形式銷售，隨著人們的數位設備儲存容量擴增以及網絡連接益加可靠，消費者開始願意將他們購買的歌曲保存在網絡上，並串流傳輸至手機上聆聽。除此之外，唱片公司也願意嘗試新的銷售模式，以歌曲在平台上的串流播放次數收費，而不是一次性地賣斷歌曲給消費者。

1 Leandro DalleMule 和 Thomas H. Davenport，〈您的資料策略是什麼？〉（What's Your Data Strategy?），《Harvard Business Review》，2017年5月至6月：112–121（*https://oreil.ly/kBC23*）。

現在世界上最大的音樂串流服務 Spotify 就是這樣開始的。可以說，Spotify 的存在正歸功於資料治理。由於當時盜版音樂正在摧毀整個音樂產業，於是 Spotify 最初以數位版權管理保護的音樂為主要業務，並提供唱片公司從他們的音樂作品中獲得報酬的一種方式。Spotify 的完整商業模式建立在圍繞追蹤用戶播放的歌曲，並因這些歌曲向音樂人回饋有償報酬。這樣的商業模式能夠證明其對資料的處理值得信賴，這是 Spotify 成為可行的音樂服務首要原因。

Spotify 密切關注使用者播放的歌曲，這項事實表示它擁有大家都在聽些什麼歌曲的資料。因此，Spotify 現在可以向聽眾推薦新歌。此類推薦演算法主要基於以下 3 點：

- 查找您所聽的音樂人其他歌曲，或相近風格的其他歌曲（例如 1940 年代爵士樂）。這稱為**基於內容的推薦**。
- 找到與您有相近喜好的其他使用者，並將這些使用者喜歡的歌曲推薦給您。這稱為**協同過濾**。
- 使用模型分析您喜歡歌曲的原始音檔，並推薦相似的歌曲。原始音檔能捕捉許多固有特徵，例如節拍。如果您偏好節奏快、樂句重複的音樂，該演算法可以推薦結構相似的其他歌曲。這稱為**相似度匹配**。

Spotify 的工程師愛德華．紐特（Edward Newett）在當時有一個有趣的想法[2]：與其一次只推薦一首歌曲，何不如讓 Spotify 創建推薦的播放列表？因此，每個星期一，Spotify 都會為每個使用者推薦一些他可能喜歡的音樂曲目，即是「個人化音樂播放清單」。

「個人化音樂播放清單」大獲成功，推出後一年內，超過 4000 萬人使用該服務，並播放近 50 億首曲目。深度個性化已經奏效。「個人化音樂播放清單」的音樂讓使用者聽起來很熟悉，但仍然保有新奇感。這項服務讓音樂愛好者可以發現新作品，讓新音樂人找到聽眾，它還為 Spotify 的客戶提供每週都值得期待的活動。

2 *https://oreil.ly/3ZQnQ.*

Spotify 能夠使用其推薦演算法，為音樂人和唱片公司提供相關粉絲偏好的見解；也可以利用推薦演算法來擴展使用者的音樂偏好，並介紹新興樂團。這種額外的知識和行銷能力，讓 Spotify 能夠在與音樂發行商的談判中占據優勢地位。

關於使用者的聆聽資訊習慣，如果 Spotify 沒有辦法保證會以負責任的方式用於改善使用者自身音樂聆聽體驗，則這一切都不可能實現。有鑒於歐洲監管機構非常保護歐盟公民的隱私，總部位於歐洲的 Spotify 如果無法證明它擁有強大的隱私控制舉措，並且確保資料科學家可以兼顧設計演算法而不會洩露使用者個人資料，就不可能讓其推薦系統落地成真。

「個人化音樂播放清單」展示了適當管理下的資料，如何創造出深受喜愛的品牌，並改變整個行業的市場領先者地位。除此之外，Spotify 藉由「年度回顧」（Spotify Wrapped）功能，擴展它的推薦系統影響力，世界各地的聽眾都可以在這個功能上，深入了解他們一年中最難忘的聆聽時刻。這是讓人們記住和分享他們最常聽的歌曲和音樂人的好方法（見圖 1-1）。

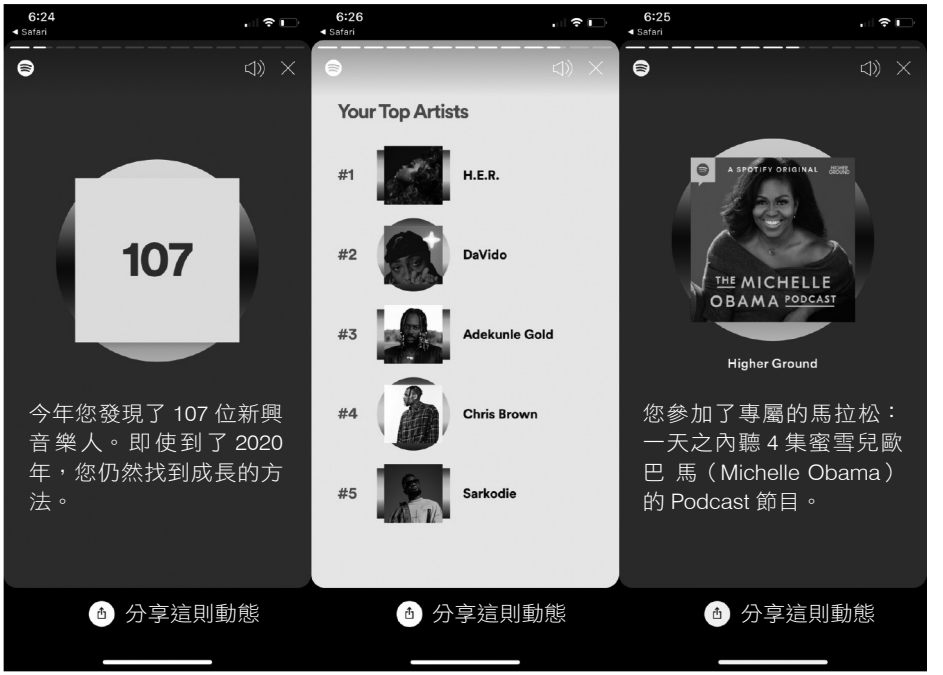

圖 1-1　本書作者之一 Anita Kibunguchy-Grant 的 2020 年度回顧播放列表

資料治理的整體方法

幾年前，當帶有 GPS 感測器的智慧型手機開始變得無所不在時，本書的其中一位作者正在研究機器學習演算法，以預測冰雹發生。然而機器學習需要標記資料才能訓練，但基於研究團隊所需的時間和空間解析度條件，非常缺乏這些資料，我們的團隊於是想到創建一個行動應用程式的想法，該應用程式允許公民科學家（非專業科學家）回報他們所在位置的冰雹資訊[3]。這是我們第一次可以自行選擇所蒐集的資料；在此之前，不管國家氣象局（National Weather Service）蒐集哪些資料，我們都只能照單全收。考慮到學術環境中，資訊安全工具處於早期且未完全發展成熟的狀態，我們決定放棄回報資料內的所有個人身分資訊，並使其完全匿名，即使這意味著報告內的某些資訊類型變得有些不可靠。但從另一個角度來看，這些匿名資料也帶來不少好處：我們開始以更高解析度來評估冰雹演算法，這樣的方式進一步地提高預測品質。這個新資料集能夠校準現有資料集，從而也提高其他資料集的資料品質。這些延伸好處不只帶來資料品質，並開始增加可信度；由於公民科學家的參與是非常新穎的點子，因此，國家公共廣播電台（National Public Radio, NPR）報導了該專案，並強調匿名資料蒐集所帶來的影響力[4]。從資料治理的角度回顧審視，它讓我們仔細思考應該蒐集哪些資料回報，改善企業資料的品質，強化國家氣象局的預報品質，甚至有助於增強氣象產業的整體形象。這種合規性、更好的資料品質、新的商業機會和信賴度強化的效果組合，是資料治理整體方法的結果。

經過幾年之後，現在我們都是 Google 公有雲的工程團隊一員，並為可擴展的雲端資料倉儲和資料湖泊構建技術。我們的企業客戶反覆關注的問題之一，是他們應該採用哪些最佳實踐和策略來管理資料的分類、發現、可用性、可存取性、完整性和安全性；也就是所謂的資料治理，客戶此時此刻因面對它而感到擔憂的心情，與我們在學術界的小團隊時期一樣。

然而，企業可用於執行資料治理的工具和能力非常強大且多樣化。我們希望能說服您，不用害怕資料治理，正確地使用資料治理可以開闢全新可能性。雖然您最

3 此移動應用程式是「近地氣象現象識別」專案（mPING，*https://mping.ou.edu*），由 NSSL、奧克拉荷馬州立大學和中尺度氣象研究合作研究所（Cooperative Institute for Mesoscale Meteorological Studies）共同合作開發。

4 這是廣播報導，但您可以在 NPR 的 All Tech Considered 部落格上進一步閱讀相關資訊（*https://oreil.ly/uWwml*）。

初可能只是服從法律或以合規性的角度來處理資料治理，但應用治理策略，可以推動業務目標增長並降低成本。

增強對資料的信任

最後，資料治理的終極目標是建立起對資料的信任。資料治理之所以有價值，是因為它增加了利益相關者對資料的信任——特別是對資料蒐集、分析、發布或使用方式的信任。

確保對資料的可信賴度，需要資料治理政策以解決 3 個關鍵方面：**可發現性**、**安全性**和**當責性**（見圖 1-2）。可發現性本身即需要資料治理，以便可隨時取用關於技術面的元資料、歷程追蹤資訊和業務面的詞彙表；此外，業務面的關鍵資料需要保持正確和完整。最後，熟悉資料管理方法以確保資料能被精細分類並得到適當保護，防止資料遭到無意或惡意的更改與洩漏。在安全性、合規性方面，敏感資料如個人識別資訊的管理，針對資料安全以及資料洩露的預防措施都很重要，具體取決於業務領域和相關的資料集。如果可發現性和安全性已經到位，就可以開始將資料本身視為一種產品。在這一點上，當責性因此至關重要，並且有必要設定資料範圍邊界，以界定資料所有權，和提供符合當責性的經營方式。

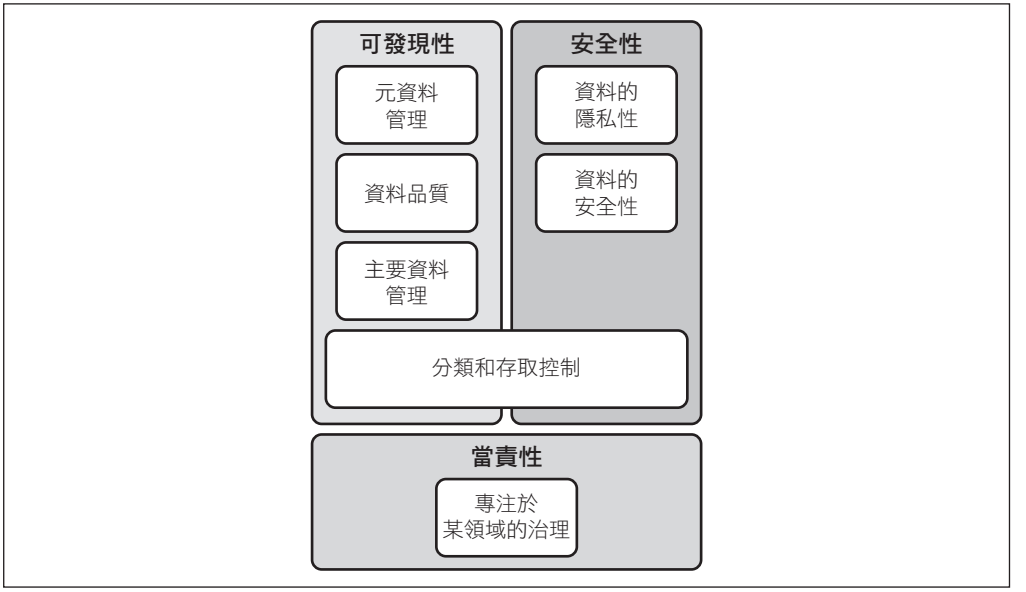

圖 1-2　資料治理必須解決的 3 個關鍵方面，以增強對資料的信任

分類和存取控制

雖然資料治理的目的是提高企業資料的可信賴度以獲取商業利益，但與資料治理相關的主要活動，仍然會涉及如何分類和相關的存取控制。因此，要了解資料治理中涉及的每一個角色，考慮典型的分類方法和存取控制設定會很有幫助。

以保護員工人事資訊為例，如圖 1-3 所示。

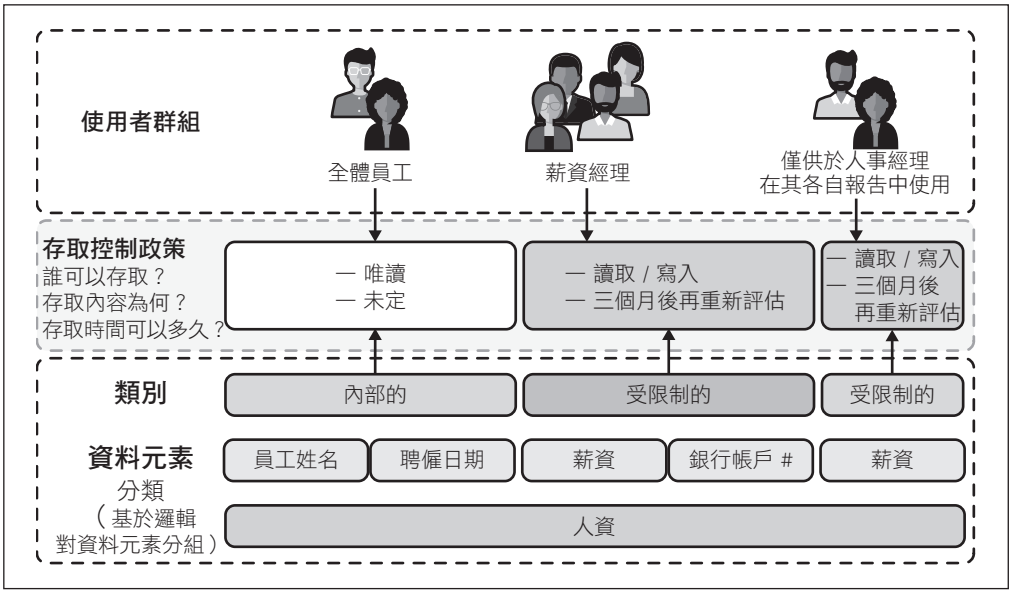

圖 1-3　保護員工的人事資訊

人事資訊包含幾項資料元素：員工的姓名、聘僱日期、過往薪資、現在薪資和受薪帳戶等。這些資料元素中的每一個欄位都以不同方式受到保護，具體取決於分類等級。而資料預設的分類等級可能是公開的，意指與企業無關的人都可以存取的內容；外部的，意指公司的合作夥伴和供應商，並且有權限可以存取公司內部系統中可存取的內容；內部的，公司組織內的任何員工都可以存取的內容；和受到限制的，例如，只有薪資處理團隊的經理可以存取每個員工的薪資，以及存入銀行帳戶的資訊。另一方面，受限程度也可能相對靈活，例如，員工當前的薪水可能只有他們的直屬上司可以看見，而每個管理階層也可能只能看到他們各自負責的團隊薪水資訊。存取控制策略將指定使用者在存取資料時可以做的事，如是否可以創建新紀錄，或者讀取、更新與刪除現有紀錄。

治理策略通常由負責該資料的部門指定，稱為管理者，如此處的人力資源部門。策略本身可能由操作資料庫系統或應用程式的團隊實作，此處為 IT 部門，因此將使用者添加到允許存取資料的群組等設定更改，通常就由 IT 團隊執行，故該 IT 團隊的成員常稱為**審核者**或是**資料管家**。而**使用者**則是指其行為受資料治理政策所允許或限制的人。在並非所有員工都可以讀取企業資料的公司中，有權讀取的員工可通稱為**知識工作者**，以和無法讀取資料的員工區分。

有的公司對於內部資料的立場採**開放**態度，比如業務資料，公司內所有知識工作者都是獲得存取授權的使用者；也有一些公司對於內部資料的立場採**關閉**態度，比如業務資料只能授權給對那些需要知道的人使用。此外，通常是由組織中的資料治理委員會來決定這類政策，至於哪種方法比較好，則沒有唯一正確的答案。

資料治理與資料賦能和資料安全

資料治理通常與資料賦能、資料安全混為一談。這些主題相互交叉，但其關注的重點各有不同：

- 資料治理主要著重於建立資料索引，以正確存取資料，方便相關人員搜尋，通常指的就是整個組織的知識工作者。這是資料治理的關鍵部分，需要「元資料索引」、用以選購所需資料的「資料目錄」等諸如此類的工具來協助。除此之外，資料治理還進一步將資料賦能擴展至資料採集階段的**工作流**。使用者可以藉由情境和描述來搜尋資料，找到相關資料的存儲位置，並附上想要的使用方式作為理由以請求存取權限。審核者（資料管家）將為之審查，確保該請求是否合理，以及所請求的資料是否可以實際應用於所提出的使用案例，如果一切條件都滿足，審核者就會啟動工作流程以允許存取資料。
- 資料賦能不僅僅是讓資料具備可存取性和可發現性，它還將資料治理擴展為允許快速分析和處理資料以回答與業務相關問題的工具：「業務在這個目標上花了多少時間和費用？」、「我們能夠優化這個供應鏈嗎？」等等的問題。該主題至關重要，需要了解如何使用資料以及**資料的實際含義**，最好的解決方法是從一開始的資料蒐集階段，就須包括描述資料的元資料，包括其價值主張、資料來源、資料歷程和對應的聯繫人知曉誰擁有、管理此資料，以便進一步查詢。

- 至於資料安全，通常可認為是一套用於防止和阻止未經授權存取的機制，因此，它與資料賦能、資料治理等方面都互有交集。資料治理依賴於到位的資料安全機制，但不僅僅只是防止未經授權的存取，還涉及有關資料本身的策略、根據資料類別以進行的轉換（參見第七章），以及證明隨著時間的推移，遵守存取和轉換資料政策的能力。總結來說，正確實施資料安全機制可以促進資料的可信賴性，使資料共享更加地廣泛或「民主化地存取」資料。

為什麼資料治理越來越重要？

自從需要治理資料以來，資料治理就一直存在，儘管它通常僅限於受監管行業的 IT 部門，以及圍繞著特定的資料集如身分驗證憑證的安全問題。即便如此，就算是舊有的資料處理系統也需要一種方法，不僅可以確保資料品質，還可以控制對資料的存取。

傳統上，資料治理向來視為一個單獨的 IT 功能，並在與資料來源類型相關的資料孤島中執行。例如，一家公司的人力資源資料和財務資料，通常是高度受控的資料，會具有嚴格控制的存取權限和特定的使用指南，而控制它們的是一個 IT 孤島；而銷售資料則位於另一個限制程度較少的孤島中。某些組織可能以「整體」的角度或「集中」的方式，來執行資料治理，但大多數公司將資料治理視為各個部門所關注的問題。

由於最近引入了 GDPR[5] 和 CCPA 類型 [6] 法規，使得資料治理成為一門顯學，影響到每個行業，不僅僅是醫療保健、金融和其他一些受監管的行業。越來越多人意識到資料的商業價值。正因為如此，現今大家看待資料治理的方式與以往相比已經大不相同。

以下只是談論資料領域的樣貌隨著時間變化的幾種面向，並且，基於這些變化，我們會將進一步討論對應的資料治理作法。

5 *https://gdpr.eu/what-is-gdpr*

6 *https://oag.ca.gov/privacy/ccpa*

資料量正在增長

現在可以蒐集的資料種類和數量幾乎沒有限制。國際資料公司（International Data Corporation）[7] 於 2018 年 11 月發布的白皮書中，預測到 2025 年時，全球的資料領域總和將激增至 175 ZB（見圖 1-4）[8]。

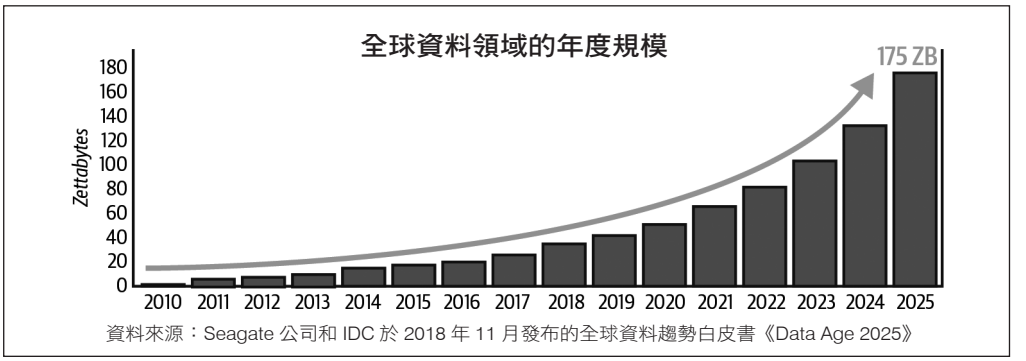

圖 1-4　全球資料領域總量規模預計將呈現急劇增長

隨著使用各式各樣新穎的技術以捕捉到更多的使用者資料，再加上預測分析，以致系統對當今使用者的了解幾乎超過使用者本身的自我認識。

處理和查看資料的人數呈現指數級增長

Indeed 人力資源網站的一份報告顯示，2015 年至 2018 年之間，市場對資料科學工作的需求驟增 78%[9]。國際資料公司（IDC）的報告也提及到目前為止，世界上有超過 50 億人正在與資料互動中，並預計 2025 年時，這個數字會增加至 60 億，占世界人口近 75%。有鑑於此，許多公司都迫切希望能夠做出「資料驅動的決策」，但這需要大量員工：從設置資料渠道的工程師，到負責資料管理和分析的分析師，再到查看儀表板和報告的業務利益相關者。當參與這份工作和查看資

7　*https://www.idc.com/about*

8　David Reinsel、John Gantz 和 John Rydning，（暫譯）《數位化的世界：從邊緣到核心》（The Digitization of the World: From Edge to Core），2018 年 11 月（*https://oreil.ly/2L1TW*）。

9　〈2019 美國最佳年度工作排名〉（The Best Jobs in the US: 2019），Indeed 人力資源網站，2019 年 3 月 19 日（*https://oreil.ly/UpU9N/*）。

料的人越多，就越需要複雜的系統來管理資料存取、處理和使用，因為濫用資料的可能性也會隨之增加。

資料蒐集方法有進步

當公司需要分析資料時，不再僅限於使用批次處理和線下處理方式，現在，公司可以利用即時或近乎即時的串流資料和分析，來為客戶提供更好、更個性化的服務與互動。客戶現在希望無論身在何處，都能透過任何連接，在任何設備上存取產品和服務。國際資料公司（IDC）預測，到 2025 年，注入業務工作流程和個人生活流程的資料，將導致全球近 30% 的資料領域成為即時資料，如圖 1-5 所示[10]。

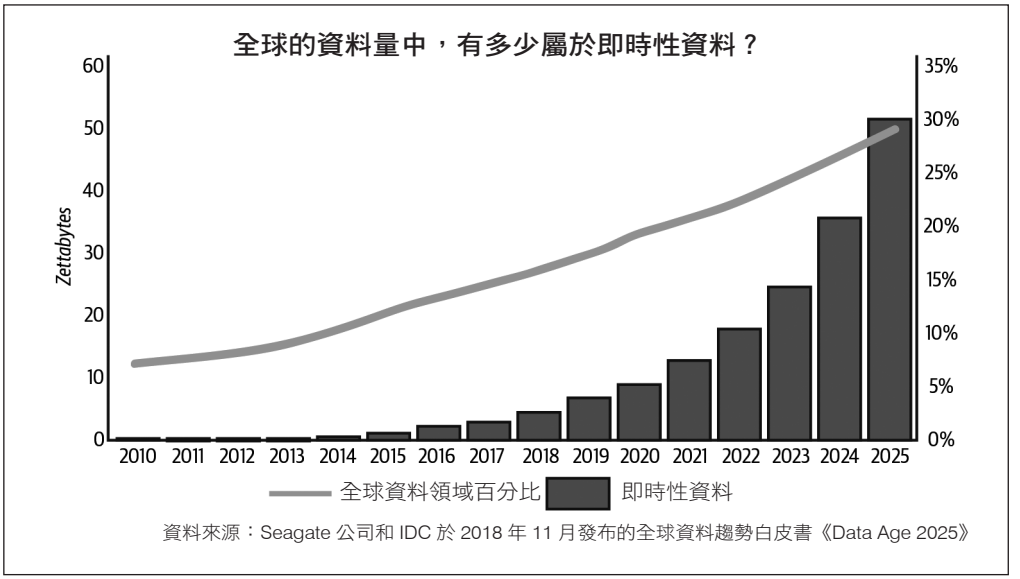

圖 1-5　到 2025 年，全球資料領域將有超過 25% 的資料是即時資料

串流技術的出現在大大提高分析速度的同時，也帶來了潛在的滲透風險，需要以複雜的設置和監控方式保護資料。

10　Reinsel 等人，（暫譯）《數位化的世界》（The Digitization of the World）。

運動賽事中的先進資料蒐集

過去，當您討論賽事統計時，您說的是相對上較為粗略的資料，比如輸贏。在某些運動中，您可能掌握球員表現的資訊，例如板球運動員每局的平均得分。然而，現在體育運動蒐集的資料，不論數量和類型都有巨大變化，因為相關團隊希望更佳了解他們可以利用的籌碼有哪些，以便在這項競爭激烈的領域中取得戰績。

因此，美國美式足球聯盟（NFL）希望更有效量化比賽中的表現也就不足為奇了，這就是它在 2015 年開始對全聯盟球隊量化分析的原因。如果您不熟悉美式足球的話，只需要記得這是一項複雜的運動，主要由美國美式足球聯盟（NFL）管理，這個美式足球的職業聯盟由 32 支球隊組成，平均分為國家橄欖球聯會（NFC）和美國美式足球聯會（AFC）。

「每次持球前進碼數」或「總衝碼數」等傳統指標可能有些缺點；當美國美式足球聯盟（NFL）了解需要進一步發展其分析和資料蒐集過程之後，便創建了次世代統計方法（NGS），這是一項聯賽專案，每個球員和裁判穿的內部襯墊，以及比賽用球、橘色方柱和衡量第一檔進攻所使用的鏈條，都貼有無線射頻識別（RFID）晶片，這個技術會讓賽事人員在每場比賽結束後得到一組非常可靠的資料統計。而正是這些統計資料構成每場比賽中每位球員在場上任一角落，包括位置、速率、速度和加速度的即時資料（見圖 1-6）。

圖 1-6　Kaggle 賽事中 NFL 資料的分析和視覺化範例。Rob Mulla 的筆記 [11]

NFL 希望得到答案的問題類型包括，「是什麼原因讓比賽中的進攻成功？」它想知道之所以成功，是取決於持球者的空手時間，還是取決於隊友的阻擋，或是教練的戰術？甚至，資料可以顯示防守方扮演的角色及其採取的行

11　*https://oreil.ly/E_XRx*

動嗎？ NFL 還希望預測一支球隊在給定進攻次數中可以前進多少碼。對進攻的深入洞察最終有助於球隊、媒體和球迷更能理解球員的技能組合和教練的策略。

正因為如此，聯盟每年都會舉辦 NFL 的巨量資料獎杯賽，這是一項體育賽事分析競賽，旨在挑戰資料分析界從大學生到專業人士的才華橫溢成員，他們為 NFL 使用進階資料分析方面的持續發展做出貢獻[12]。參賽者需分析和重新思考趨勢及球員表現，讓美式足球的比賽和訓練方式得到創新。

以上這個 NFL 範例展示了資料蒐集方法的進步程度。毫不意外地，這才是真正加速全球資料領域生成更多資料量的原因。

世界正在蒐集更多種類的資料（包括相較敏感的資料）

預計到 2025 年，每人每天藉著與數位服務的互動，將會進而創建超過 4,900 次資料；大約每 18 秒就會有一次數位資料創建（見圖 1-7）[13]。

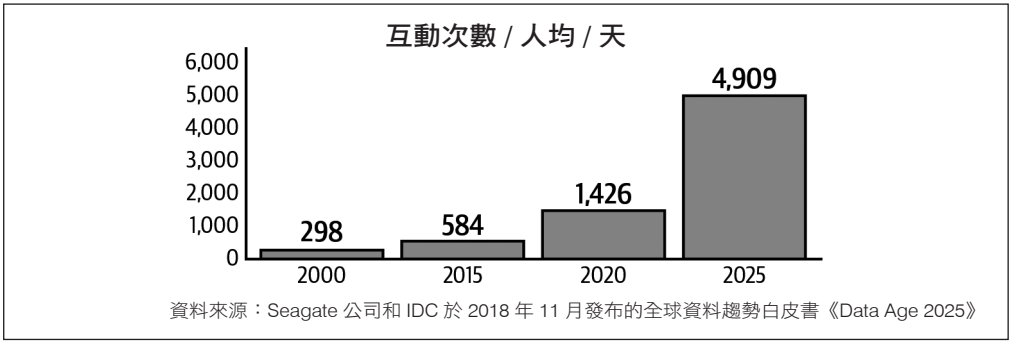

圖 1-7　到 2025 年，每人每天藉著與數位服務的互動，進而創建超過 4,900 次資料

其中許多互動將產生和蒐集大量敏感資料，例如個人身分證號碼、信用卡號碼、姓名、地址和健康狀況等。這些極其敏感類型的資料蒐集激增，對如何使用和處理這些資料，以及誰可以查看，都會引起該服務使用者和監管機構的極大關注。

12　Kaggle：NFL 巨量資料獎杯賽（NFL Big Data Bowl，*https://oreil.ly/o7lCI*）。

13　Reinsel 等人，（暫譯）《數位化的世界》（The Digitization of the World）。