名家專欄

下一代資料平台是資料網格(上)

口述／孫培然　彙整／CIO編輯室

上一期說明醫院在數位轉型時，會面臨的主要問題之一，就是傳統式資料平台形成的瓶頸。為了解決這個問題，醫院可能需要導入以資料網格（Data Mesh）為基礎的現代化資料平台。

資料網格的定義

資料網格是一種基於領域導向（Domain-oriented）和自助服務資料平台的新模式，借鑒了微服務的分散式架構思想，最初源於ThoughtWorks首席技術顧問 ZhamakDehghani 發表在MartinFowler 官網上的兩篇文章「How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh」和「Data Mesh Principles and Logical Architecture」。討論著如何從一個耦合度極高的中心化資料湖（Data Lake），轉移到分散式的資料網格。在 2019 年時，ZhamakDehghani 提出了資料網格的方法論，主要聚焦在四個大原則，如下所述。

第一個原則是以領域導向的資料所有權（Domain Ownership），也就是資料本來的所有權是在大數據中心或者資訊室，現在把所有權下放到各科領域，如神經外科有屬於自己責任範圍要管理的資料，同時也可使用自己的資料去產生更多資料。同理可證，腎臟科也一樣，各科有專門管理資料的人員，而且是以低代碼/無代碼（Low-code／No-code）的概念去完成資料治理。

第二個原則是要把資料當成產品（Data as a product ），例如神經內科可能需要 15 個或 30 個欄位時，不再只是依照不同的需求提供不同的欄位，而是神經內科應該包括哪些資料，就把它打包成一個產品提供給使用者，也就是資料即產品的概念。

Vision

第三個原則是自助資料基礎架構平台的管理（Self-serve data infrastructure platform），也就是了解領域專業的人，其實不用寫程式，而是只要透過視覺化平台的工具，去設定來源資料與目的資料的映射（Mapping）去做想要做的事情，便可以將來源資料根據你的設定，即時的轉到你所指定目的資料中台上，提供使用者存取使用及研究分析。

第四個原則是聯合計算治理（Federated computational governance），因為資料所有權已經下放，為了避免形成資料孤島，就要透過一個聯合計算的治理過程，站在使用者角度去了解資料如何被應用及分析，制定資料治理相關規範，讓資料更為透明和易於使用。

‖ 以領域導向的資料所有權

以領域導向的資料所有權，主要概念就是從業務領域角度出發，將本來集中在大數據中心的資料去中心化，將業務解耦之後，映射到資料的視角，再將資料解耦，減少資料冗餘度，重新考慮整個資料的處理位置和所有權。

再來就是臨床作業，強調的是以「就源輸入」，也就是資料在哪裡產生，就在哪裡負責輸入的概念，看誰有能力能夠了解資料的意義，以及有辦法進行資料解構或重組，那就由誰來負責。

最後就是如何使用這些資料，以及改變現有資料產生及使用的方式，誰最清楚就由誰來負責，也就是以領域知識（Domain Knowledge）的概念，把所有權下放到使用者或者生產者的概念。

‖ 以資料即產品的思維

資料即產品的思維，主要是提出了一個「資料產品經理」的概念。其職責包括，如何讓使用者可以很輕鬆隨心所欲的存取跟應用資料，讓資料可以更容易的去取得。

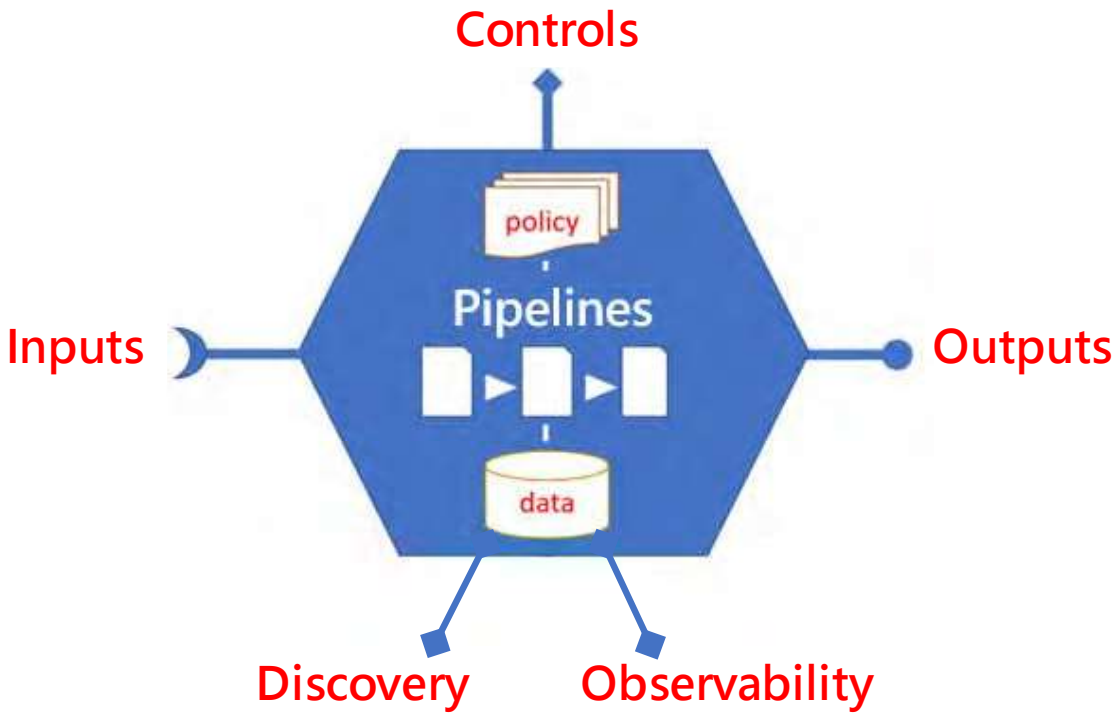
為了讓資料可以更容易取得，就必須要建構出完整的資料服務目錄，讓使用者可以隨選查詢他所要了解的資料，並不斷的改善、改善、再改善，改善成一個真正的產品，讓大數據工程師或IT人員可以專心的投入應用程式的相關應用，而不需要花費額外的時間去管理基礎架構。

資料即產品的思維，也可讓生產者跟使用者洞悉資料產生和使用的行為，進而鼓勵正向的使用資料行為。透過不斷的修正那些不符合資料平台思想的行為，不斷提高使用率，可以更進一步地引進出更好的資料產品設計概念。

‖ 從資料產品到架構量子

資料即產品的思維，引發出一種新的概念為「架構量子（Architectural Quantum）」是以最小的架構單元，它可以獨立部署，具有高內聚力、低耦合力，並包含其功能所需的所有結構元素，也就是將整個資料產品變成以架構量子為最小的架構單元，所以微服務就是最小的單體概念，一個典型的產品架構量子，如圖一所示，主要有六個組件：第一是要有一組資料輸入端口（Inputs），主要是接收其來源資料；第二是要有一組連接控制

專欄專家孫培然博士，私立醫療院所協會醫院資訊暨智慧醫療發展促進會會長，也任職於中國醫藥大學附設醫院資訊室副主任。



圖一、架構量子

介面（Controls），可以在其中控制資料載入及資料儲存的相關政策，以確保資料的品質；第三是要有一組內置的資料流水線（Pipelines），如資料進來時，會經過很多不同的流水線，比如西元年會轉成民國，再輸出給需要的人；第四是要有可發現介面（Discovery），讓使用者很輕鬆容易的找到他們所需要的資料描述，包含元資料（Metadata）及存取資料所有權等；第五是可觀察性介面（Observability），主要是幫助使用者了解量子中的資料品質，並確保資料是否符合他們的服務水準目標（Service-Level Objective）；最後，第六要有一組資料輸出端口（Outputs），提供一個可以互交操作輸出模式，其模式可以是資料庫、文件、API等。

‖ 資料即產品的特性

資料即產品的特色，第一個是「可發現」，也就是透過資料目錄，可以去包括所有權和內容的元資料，有助於使用者找到需要的資料；第二個是「可定址」，也就是每個可發現的產品都有一個唯一識別碼，以便可以對其進行定址存取；第三個是「自我描述」，資料產品必須要為其預期的使用者提供清晰的語義、語法跟資料庫的模式；第四個是「可信賴」，對領域資料的所有者制定服務級別目標，管理其資料產品的可信度；第五個是「安全」針對資料存取權限的管控機制，確保資料的安全被存取；最後是「可交互」，資料網格中的資料產品應該要能夠跨領域存取。

要做到資料即產品，必須要有自主資料服務工具，並且要把它抽象化到與領域無關，不然就是等於客製化，日後就會形成一個資訊孤島，所以與領域無關的自助服務工具，要能夠讓資料基

圖二、跨服務、系統的多路雙向資料發布管線

礎建設平台，脫離中心化的領域資料所有權。

我們要提供一個與業務領域無關的資料基礎建設平台，透過這個平台可以很迅速的將很多資料轉換成資料產品，讓它可以擁有高內聚力、低耦合力，進而成為一個能夠相互溝通有無的資料產品工具平台服務。

自助資料服務基礎建設平台，要做到持續整合及持續交付（CI/CD）貫穿於應用的整個生命週期，才能形成很多個資料產品，也就是架構量子。它們之間則是透過 API 來互相溝通，也就是資料網格的概念。整個框架平台就等於是整個資料網格的生態系，可以具有相關的領域特色跟領域專業。自助資料服務基礎建設平台，就可以具資料產品思維的特性，來做管理跟角色的區別，做到持續整合、持續建置的概念。

每一個架構量子之間，可以透過建立資料管線及設定檔做資料轉換，這就是最近常在講的「Low-code」及「No-code」的概念。也就是架構量子必須要可以讓資料所有權者，可以去做資料擷取、轉換跟輸出的行為，讓使用者可以自定義一個資料中台。

自助資料服務平台的基礎建設，大概包括幾個功能，如資料的快取跟快照、資料來源的描述與原始副本、資料產品範本、資料格式轉換、資料管線範本、資料存取控制，以及日誌與追蹤。

以圖二為例，比如說來源資料庫可能會是 RDBMS、Hadoop或是 NoSQL，假設這些資料是某專業領域所需的資料，就可以透過資料服務平台的基礎建設，在設定完成以後，資料來源就會自動映射到目的地資料庫，也就是使用者所要使用的資料庫，中間就是要把架構量子相關資料管線及設定檔，讓使用者去設定相關的轉置機制，也就是資料互相轉換跟交換的概念。