資料整理檢查清單

歡迎來到人工智慧 (AI) 的時代，在這裡您需要仰賴資料密集型科技(例如機器學習與深度學習)來營運。若要充分利用這些新的 AI 工具，您必須確認 組織的資料「空間」井然有序。

這是一份您著手進行資料整理的待辦事項，其中將 AI 的關鍵階段分為兩種：訓練與推論。

遵循下列步驟有助於您成為 AI 專家。若要進一步了解如何將 AI 從概念驗證到全面量產與規模化，請參照下列 IDC 報告，*Accelerate and Operationalize AI Deployments Using AI-Optimized Infrastructure (使用 AI 優化的基礎架構加速並落實 AI 部署).*

訓練

在準備 AI 的訓練方面，您將會研發演算法，以對資料集有所瞭解。您的主要重點將是蒐集現有資料並使用 AI 來學習新的功能。

- [ ] 利用 AI找出您希望解決的特定業務問題(從較小的專案開始，有助提升您的學習效果)
- [ ] 找出可解決相關問題的資料(大部分的資料都不會位於單一位置)
- [ ] 利用資料標籤來準備資料，以大幅縮短尋找相關資料所需的時間
- [ ] 在您將使用的所有資料集之間，確認您的資料已適當同步與連結(包含同步的時間)
- [ ] 為任何敏感客戶與其他個資加註旗標，裨使您能確保資料安全無虞，而且可遵守所有 適當的法律與規約(元資料標示過程有助於此作業)
- [ ] 為您正在使用的資料類型選擇適當的開發環境，以及將進行格式化的方式(例如：圖片、影片、格式不拘 的文字與音檔 - 每種格式通常都屬於一種環境)
- [ ] 從儲存庫拉出資料集並將其導入您的開發環境
- [ ] 將您的資料分成兩個群組，以協助改善您的模型制定流程(將一個資料集保留在名稱為「train」的資料夾，另一個資料集則保留在名稱為「test」的資料夾)
- [ ] 追蹤資料位置/來源以維持資料的可追溯性(考慮使用有助於自動化此過程的工具)
- [ ] 執行基本資料清理任務，以準備建造模型的資料(例如：填入遺失的資料項目並移除空白的項目)
- [ ] 使用一組您已經掌握答案的預測活動作為資料子集範例(名為「訓練集」)，並且在進行預測前確認所有資料前處理步驟都已到位
- [ ] 運用您對這個訓練集的了解來給予恰當分數，如此您就能自信地將同一模型套用到尚未被明確訓練模型所使用過的新資料上

推論

一旦您完成可解決業務問題的模型開發，您就能從訓練階段進展到到推論階段。在此階段，您使用成功的模型並將之套用到新的資料，這也需要持續進行一些資料整理作業。

- [ ] 將您的 AI 模型放在接近資料的位置，以減少延遲、減少頻寬要求以及增進整體模型效能
- [ ] 建立有效的資料通道，並在資料傳入時將元資料標籤套用至您的資料中，如此就能蒐集新資料並將之用於 強化未來的模型運轉
- [ ] 以連結及同步化的方式標示資料(例如，如果資料是依照時間順序排列的，您可以 跨資料集進行同步，或在所有傳入的資料上選擇一個欄位進行連結(例如，客戶的名稱)
- [ ] 制定長期的儲存資料生命週期計畫，以了解如何在資料傳入與儲存時管理資料的數量及速度
- [ ] 考慮招募資料長，為未來的 AI、深度學習與其他資料驅動專案維持組織的資料管理

IBM