數位無限軟體股份有限公司
(07) 396 3396 | sales@infinitiessoft.com
www.infinitiessoft.com | www.facebook.com/InfinitiesSoft

Accelerate Muitl-Cloud AI computing with InfinitiesSoft AI-STACK.
Contact our cloud experts today for a free consult!

兩種AI-STACK方案套裝
滿足不同階段的需求

Let us help with your!

以AI-STACK LITE 搭配GPU伺服器
簡便使用提高工作效率

- 提升系統與用戶資源可視可控的管理能力
- 實現多用戶單機資源共享(用戶運算與本機硬碟環境隔離)
- 容器模板與資源規格選單化自動供裝提升工作效率

AI-STACK EXPRESS 搭配多台GPU伺服器
組建POD協同管理

- 可結合用戶身份認證系統並可掛載既有儲存資源
- 資源共享更安全(支援容器IP白名單+儲存隔離)
- 個人和團隊更高效協作的自動化、批次與排程功能

AI-STACK 滿足
IT管理者與機器學習工作者需要

1. **降低使用者門檻**
   簡易步驟快速建立
   (個人/團隊) ML環境
2. **掌握有效資源**
   彈性IT資源共享、額
   度限制、工作排程
3. **環境自主權**
   IT環境隔離、SSH Key
   或密碼登入環境
4. **易於使用和共用**
   帳號/儲存整合,批次、
   預約申請作業

產品功能介紹

壹、AI-Stack核心架構

1. 提供Web操作介面，以圖形化操作方式進行深度學習容器自助式申請與建立容器操作。
2. 使用Kubernetes進行容器調度，能對不同使用者進行資源自動分配與部署。
3. 平台內建NVIDIA 優化之常用TensorFlow、PyTorch之 AI 框架，並具備 AI 框架擴充設計。
4. 在軟體授權範圍內可納管一台或多台AI運算節點，成為一個或多個用戶共享資源池。
5. 內建符合OMG開放性流程引擎的BPMN功能(Business Process Model and Notation，業務流程模型和標記法)，提升平台服務自動化的建立、調整及管理。
6. 平台可介接外部存儲設備，可透過NFS介接標準NAS。
7. 平台可介接LDAP/AD、OpenID、OAuth等用戶身份登入認證機制。
8. 平台支援多容器共享GPU機制，讓多名使用者、多個容器可共用同一片GPU進行操作開發。
9. 平台支援NVIDIA A100之MIG(Multi Instance GPU)架構，可於Web操作介面使用MIG規格

貳、使用者操作介面

1. 平台提供線上機器學習服務申裝模式，讓使用者採用自主服務的方式申裝、管理運算資源，內容含有：
   (1) 使用者可自行選擇資源規格，依需選取所需的GPU張數、CPU core數量、Memory數量。
   (2) 使用者可自行選擇含AI Framework(例如：Tensorflow)之容器範本。
   (3) 使用者自助申請建立、刪除容器服務。
   (4) 使用者可預約容器自動建立時間並於時間結束時自動釋放容器資源功能。
   (5) 使用者可細部查詢容器資訊，平台透過列表方式呈現包含機器學習服務名稱、硬體配置、所選擇的Framework、部署 ID (UUID)、機器學習服務建立時間、機器學習服務的擁有者、目前機器學習服務的運行狀態。
   (6) 此平台提供 SSH Key 的SSH連線容器登入方式或提供輸入密碼的SSH連線容器登入方式。
   (7) 使用者可以查看個人平台的操作記錄，包含 IP 資訊與使用者開通、登入、登出、建立AI容器、刪除AI容器之操作行為紀錄。
2. 容器建立後平台提供使用者Jupyter Notebook程式編輯工具進行開發。
3. 使用者可以針對個別容器設定連入IP白名單，僅有允許的IP可連入容器進行操作，提升網路連線安全。
4. 容器內建安全殼層 (SSH) 通訊協定，並搭配金鑰對與自訂義密碼進行安全連線。
5. 平台可產生公私金鑰對Key Pair，使用者也可自行匯入公鑰，金鑰用於容器SSH連線使用。
6. 租戶管理者可批次建立容器，即使用同一個容器樣板一次為多位相同租戶使用者建立容器，減少相同使用者個別建立與操作時間。
7. 提供Batch Job深度學習服務訓練模式，使用者將訓練任務打包為Shell Script並於平台輸入執行命令，平台將自動建立容器並執行Batch Job內容，且於訓練結束時自動刪除容器。
8. 提供Batch Job排程工具，使用者可設定執行時間，平台於排程期間將定時執行重複性訓練任務。

9. 提供使用者自定義鏡像(Image)上傳與製作Template樣版之功能，使用者將可打造專屬工作環境容器，節省程式套件反覆安裝的時間與人力。
10. 提供share memory動態調整功能，使用者建立GPU資源時可自行輸入共享記憶體大小(上限70%)。
11. 於建立容器時以圖像化提示當下伺服器GPU剩餘片數，若資源不足將提醒使用者本次申請可能需要等候。

參、管理者操作介面

1. 平台提供深度學習服務維運管理，內容含有：
   (1) 平台具備容器資源規格(GPU、CPU、Memory)的管理功能，可設定資源規格的模板，供使用者選取。
   (2) 設定與修改每個租戶與每個租戶內成員的GPU、CPU、Memory資源配額，以達到資源管控目的。
   (3) 提供多租戶管理功能，管理者可依研究計劃分類，建立多個租戶，並將使用者加入租戶中形成不同群組，一個使用者可同時存在多個租戶，且可將用戶指定為租戶管理者，授予管理權限。
2. 提供儀錶板檢視功能，呈現目前平台所有伺服器節點總數量、容器總數量、GPU總數量，以及個別節點上運行的容器總數量、GPU總數量與型號，輔以圖型化呈現方式，提供管理者清晰明確的系統資訊。
3. 承上，GPU監控功能包含GPU使用率、GPU Memory使用率以及GPU溫度，並以圖像化方式顯示監控資訊。
4. 提供歷史GPU使用率與使用時數查詢功能，協助管理者清查容器使用狀態，提高GPU使用效率。
5. 提供服務成本分析圖表，管理者可於後台定義AI資源服務項目與定價，平台自動收集資源使用狀況並以圖表方式呈現用量與費用資訊。
6. 平台具備簽核與審查機制，AI資源申請可設定為需經申請人主管(租戶管理者)審核同意後建立，但未大於特定額度可不需經審核自動建立。
7. 提供後台管理者容器時限管理機制，可設定容器結束時間，增加管理彈性。
8. 提供後台管理者GPU使用率低落回收機制，當GPU使用率持續一段時間為0%時，系統將自動回收該容器，該時間長度可由管理者自訂，並於刪除前後均寄信通知使用者。