附件一：「政府巨量資料技術工具研發計畫」徵求主題細部說明

I. 巨量資料擷取與管理技術工具

巨量資料之前處理對於後端之資料分析應用極為重要，前處理主要包含資料之擷取(Extraction)、轉換 (Transformation) 與載入 (Loading) 等三大步驟，從擷取面來看巨量資料中強調的資料流處理、大量非結構化文字處理以及各種格式間之快速轉換技術工具為目前仍較缺乏部分。同時，一個能整合各種前處理模組以發揮串聯各模組功能之友善界面平台於政府巨量資料之應用上亦極為需要。以下為相關徵求研究主題:

1. 資料流前處理技術 (Streaming Data Preprocessing)

目前資料分析系統對於即時資料流的處理仍侷限在靜態資料的處理方式，使用者必須自行在時間軸上進行資料流處理，增加了分析程式撰寫上之複雜度。如何發展能針對即時資料流的資料介接與前處理界面，可以讓後端分析程式不需考慮資料源特性與時間特性，在前處理時即能對即時資料流併行分析為重要之議題。

2. 非結構化文字資料擷取與前處理技術 (Unstructured Text Data Preprocessing)

現有資料處理系統中對於非結構化文字資料，多僅提供字串處理功能，在文字上需要開發泛用之擷取與前處理功能，例如自動網頁資料擷取以及斷字與詞頻統計等，或是加入語法結構與詞性分析等，並結合現有相關文字處理工具，提供使用者透過此文字處理模組可直接將文字資料源直接處理成後續分析所需要的結構資料。

3. 快速與自動資料格式轉換技術 (Fast and Automatic Transformation of Data Formats)

同時面對不同資料格式並一起進行分析時的快速與自動格式轉換相當重要，如資料來源是網頁內容中特定且固定的資訊欄位，自動截取資訊並轉換成 JSON 格式進行資料分類、轉換與儲存，在面對未處理的網頁形式開放資料為一重要議題。

4. 快速及具可適性之前處理模組 (Fast and Adaptive Data Preprocessing Module)

現存資料前處理演算法在處理巨量資料時常有記憶體消耗與運行時間過長的問題，多數演算法在巨量資料平台上需另予最佳化改寫，如何達到快速及具備資源可適性為主要挑戰。這部分雖多屬學術研究領域，目前亦逐漸出現具體之系統開發，為一重要課題。

5. 巨量資料特徵標註技術 (Big Data Feature Labeling)

資料前處理中特徵的選取相當重要，除了可利用資訊分析的方式選取有效資料特徵外，能夠有效地整理並提供資料特徵標註的模組對於資料分析上都相當有幫助，提供使用者在資料分析前能利用這些工具選擇並標註有用或需要之特徵資料，為一重要議題。

6. 整合性架構資料前處理工具 (Integrated Big Data Preprocessing Framework)

上述前處理模組相當多且複雜，實務上在進行巨量資料處理與分析時，往往不僅只需要單一種前處理方法，目前在整合性的 framework 上仍鮮有工具，需要一個能夠提供友善介面之平台以整合多項前處理工具與加速處理流程。

II. 巨量資料分析技術工具

針對巨量資料之大量性 (Volume)、多樣性 (Variety)、快速成長 (Velocity) 等繁複特性，已有為數不少的資料分析技術與軟體以支援巨量資料的特性。然針對政府開放資料或所建構之相關雲端服務應用，例如自然環境之感測資料、交通雲之感測資料、醫療雲之健康照護資料與社群網路資料等，這些資料反映了更多樣性的特質及分析複雜度。針對政府巨量資料之多樣性、複雜性及相關應用情境需求等，進行統計分析技術研發，開發巨量資料共用軟體。相關徵求主題如下:

1. 時空資料分析技術 (Spatial-Temporal data analytics)

在政府目前的開放資料中，如環境感測資料、氣象資料、即時路況資料等，這些資料均同時具有時間與空間維度的重要資訊。因此，未來在政府巨量資料分析的工具中需要具備更多可有效分析巨量空間與時間資料之資料分析工具。

2. 圖形資料分析技術 (Graph data analytics)

針對政府巨量資料的多樣性，如健保資料、交通雲之即時路況與路網資料、社群網路資料等，提供可支援高多樣性及複雜性之巨量圖形分析演算法為重要之研究議題。

3. 預測最佳化技術 (Optimization of prediction tasks)

巨量資料環境中預測模型的精準度要達到最佳的預測準度，往往需要透過不斷的嘗試不同的建模演算法與參數調校，此過程極為耗時。針對給定之預測標的物，如何選取適當之建模演算法與最佳化參數達到最佳之預測精準度為一重要議題。

4. 前端裝置資料分析技術 (Front-End data analytics)

在某些資料收集的應用情境中，如交通雲之即時路況視訊監控等，這些多媒體或感測資料等需要更多前端的基本資料分析，以降低後台分析平台及資料傳輸的成本等。如何開發適用於前端裝置的資料分析工具亦為一重要研究議題。

5. 隱私保護的資料分析技術 (Privacy preserving data analytics)

政府資料中常具有敏感性資料，如健保就醫資料、戶籍資料、個人金融資料等，如何於資料分析過程中保有適度之資料隱私，使探勘出來的模式或是參數僅能允許資料提供者掌握與擁有為一重要研究議題。

6. 以 GPU 為基礎之資料分析技術 (GPU-based data analytics)

隨著資訊硬體的快速發展，以 GPU 為主之伺服器已漸趨普遍，例如國網中心已建置台灣最大之GPU 叢集運算環境Formosa 5，GPU 目前已經廣泛用於各種巨量資料之分析處理。因此,以 GPU 為基礎之資料分析技術為值得探索之議題。

7. 非結構化資料分析 (Unstructured data analytics)

政府巨量資料中存在許多非結構化資料， 如文字資料、影像視訊、即時交通視訊等。針對非結構化資料之分析，目前仍缺少有效的分析平台與工具，針對政府巨量資料的中的非結構化資料提出有效之分析技術工具為一重要課題。

III. 巨量資料視覺化呈現技術工具

巨量資料工具應用中，資料呈現技術為前端介面最重要的挑戰之一，政府巨量資料視覺化議題，不像以往只限於專家與專家之間的圖形溝通介面，如今更是政府，專家學者與民眾多方溝通的橋樑。而過往資料呈現多依賴單一作業系統上之統計模擬商業軟體所搭配的報表呈現功能來達成複雜的資料視覺化，因作業系統介接等問題，這類的資料呈現多不利於大眾化分享。如何考量到資訊傳達的互動性、有效性、即時性、直覺性、便利性，是這部分議題徵求的主要方向。因此，本計畫即徵求基於開源形式能支援開放性跨平台架構，符合 HTML5 架構之資料視覺化技術工具 (基於 R、Python、Javascript 等程式語言之函式庫)或整合性資料呈現服務，以便利各樣式的巨量資料呈現。相關徵求主題如下：

1. 適應各式前端裝置的資料呈現技術

隨著手持式及穿戴式裝置的普及，透過各式設備不同類型螢幕來存取資訊已成為十分普遍的做法。因此能適應各式前端裝置的螢幕，解決螢幕碎片化所帶來的挑戰，為一重要議題。

2. 大量資料下的互動式視覺化技術

對於大量資料之呈現，如合有效解決前後端大量資料傳輸延遲情況，以及克服大量資料圖表互動操作時 lag 等問題，為新型視覺化工具之重要議題。

3. 異質性資料的互動式視覺化技術

過往的 Web-based 資料呈現套件如 D3.js，多著重於單一資料的互動式呈現。然現今各式資料均能快速的累積，能簡易 mashup 異質資料集，呈現不同角度的 dashboard 式視覺化介面，為極具挑戰之重要議題。

4. 即時性資料的互動式視覺化技術

現今的 Web-based 視覺化工具主要透過 Javascript 將資料完整傳輸到使用者端後開始繪製圖表，對於處理資料流(streaming data)的呈現，將十分不便利。如何有效地只針對新的資料做運算，並將之與舊的資料及其視覺化圖表結合，也是新型態工具發展的方向之一。

5. 智慧性資料視覺化技術

巨量資料中樣式相當多元，常用格式如文字 json，伺服器存取 log，關聯式表格，影像等，如何減少非資訊背景使用者在處理整合資料呈現上會面臨的技術門檻，能智慧化判斷常見的資料應用情境，主動的產生相對應的基礎圖表以及相關程式碼，為一重要議題。

6. 整合性資料視覺式服務

現今已有許多提供資料分析與視覺化網站建置的開源整合服務平台, 如 CartoDB 以及 Rstudio 的 Shiny package 等可自建視覺化服務網站。然此類服務提供的功能多需再經過大幅修改調校，方能滿足特定領域的使用者需求。如何提供更方便的視覺呈現整合服務以加速政府各領域資料之有效呈現，是一重要之研究議題。