



(19)中華民國智慧財產局

(12)發明說明書公告本

(11)證書號數：TW I789392 B

(45)公告日：中華民國 112 (2023) 年 01 月 11 日

(21)申請案號：107114429

(22)申請日：中華民國 107 (2018) 年 04 月 27 日

(51)Int. Cl. : H04N19/85 (2014.01)

G06F16/00 (2019.01)

(30)優先權：2017/04/28 美國

62/491,864

(71)申請人：美商艾斯卡瓦公司 (美國) ASCAVA, INC. (US)

美國

(72)發明人：夏拉帕尼 哈莎夫丹 SHARANGPANI, HARSHVARDHAN (US)

(74)代理人：林志剛

(56)參考文獻：

TW 339495B

US 2013/0339323A1

US 2016/0188622A1

WO 2006/055587A2

WO 2016/205209A1

審查人員：賴文能

申請專利範圍項數：30 項 圖式數：19 共 244 頁

(54)名稱

藉由使用主要資料篩的資料無損縮減及對於已經使用主要資料篩來無損地縮減的資料進行多維度搜索和內容關聯的檢索

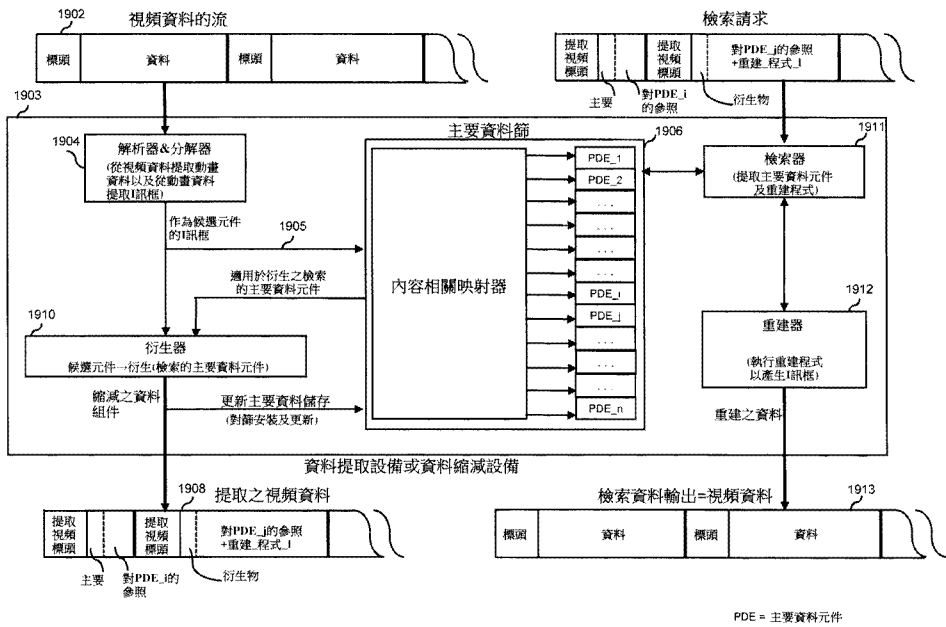
(57)摘要

藉由使用根據其內容組織主要資料元件的資料結構，可以無損縮減輸入資料。可替換地，資料結構可以根據從主要資料元件得知的名稱內容來組織主要資料元件。具體地，視頻資料可藉由(1)使用資料結構來識別一組主要資料元件，以及(2)使用該組主要資料元件來無損縮減內訊框而無損縮減。輸入資料可以根據資料結構的組件的記憶體使用率進行動態分區。可以根據分區建立包裹，以利資料的歸檔和移動。無損縮減的資料可以使用一組提取檔案和一組主要資料元件檔案來儲存。

Input data can be losslessly reduced by using a data structure that organizes prime data elements based on their contents. Alternatively, the data structure can organize prime data elements based on the contents of a name that is derived from the prime data elements. Specifically, video data can be losslessly reduced by (1) using the data structure to identify a set of prime data elements, and (2) using the set of prime data elements to losslessly reduce intra-frames. The input data can be dynamically partitioned based on the memory usage of components of the data structure. Parcels can be created based on the partitions to facilitate archiving and movement of the data. The losslessly reduced data can be stored using a set of distilled files and a set of prime data element files.

指定代表圖：

第 19 圖



符號簡單說明：

1902 . . . 視頻資料流

1903 . . . 資料提取設備

1904 . . . 解析器/分解器

1905 . . . 候選元件

1906 . . . 主要資料節

1908 . . . 提取視頻資料

1910 . . . 衍生器

1911 . . . 檢索器

1912 . . . 重建器

1913 . . . 視頻資料



I789392

【發明摘要】**【中文發明名稱】**

藉由使用主要資料篩的資料無損縮減及對於已經使用主要資料篩來無損地縮減的資料進行多維度搜索和內容關聯的檢索

【英文發明名稱】

LOSSLESS REDUCTION OF DATA BY USING A PRIME DATA SIEVE AND PERFORMING MULTIDIMENSIONAL SEARCH AND CONTENT-ASSOCIATIVE RETRIEVAL ON DATA THAT HAS BEEN LOSSLESSLY REDUCED USING A PRIME DATA SIEVE

【中文】

藉由使用根據其內容組織主要資料元件的資料結構，可以無損縮減輸入資料。可替換地，資料結構可以根據從主要資料元件得知的名稱內容來組織主要資料元件。具體地，視頻資料可藉由(1)使用資料結構來識別一組主要資料元件，以及(2)使用該組主要資料元件來無損縮減內訊框而無損縮減。輸入資料可以根據資料結構的組件的記憶體使用率進行動態分區。可以根據分區建立包裹，以利資料的歸檔和移動。無損縮減的資料可以使用一組提取檔案和一組主要資料元件檔案來儲存。

【 英文 】

Input data can be losslessly reduced by using a data structure that organizes prime data elements based on their contents. Alternatively, the data structure can organize prime data elements based on the contents of a name that is derived from the prime data elements. Specifically, video data can be losslessly reduced by (1) using the data structure to identify a set of prime data elements, and (2) using the set of prime data elements to losslessly reduce intra-frames. The input data can be dynamically partitioned based on the memory usage of components of the data structure. Parcels can be created based on the partitions to facilitate archiving and movement of the data. The losslessly reduced data can be stored using a set of distilled files and a set of prime data element files.

【指定代表圖】第(19)圖。

【代表圖之符號簡單說明】

1902：視頻資料流

1903：資料提取設備

1904：解析器/分解器

1905：候選元件

1906：主要資料篩

1908：提取視頻資料

1910：衍生器

1911：檢索器

1912：重建器

1913：視頻資料

【特徵化學式】無

【發明說明書】

【中文發明名稱】

藉由使用主要資料篩的資料無損縮減及對於已經使用主要資料篩來無損地縮減的資料進行多維度搜索和內容關聯的檢索

【英文發明名稱】

LOSSLESS REDUCTION OF DATA BY USING A PRIME DATA SIEVE AND PERFORMING MULTIDIMENSIONAL SEARCH AND CONTENT-ASSOCIATIVE RETRIEVAL ON DATA THAT HAS BEEN LOSSLESSLY REDUCED USING A PRIME DATA SIEVE

【技術領域】

【0001】此發明有關資料儲存、檢索及通訊。更具體地，此發明有關對於已經使用主要資料篩來無損縮減的資料進行多維度搜索和內容關聯的檢索。

【先前技術】

【0002】現代化的資訊時代係藉由龐大數量之資料的建立、獲得及分析為特點。新的資料係由不同的來源所產生，其範例包含採購交易記錄、企業及政府記錄和通訊、電子郵件、社交媒體文章、數位圖像和視頻、機器日誌、來自嵌入式裝置的信號、數位感測器、蜂巢式電話全球定位衛星、太空人造衛星、科學計算，以及巨大挑戰科學。

資料係以各種形式產生，且其中很大一部分係非結構化及不適合進入至傳統的資料庫之內。商業、政府及個人以空前的速率產生資料，且努力要儲存、分析及傳達此資料。每年花費數百億美元在儲存系統的採購上，以保持累積之資料。同樣地，大量金額亦被花費在電腦系統上，以供處理該等資料之用。

【0003】在最現代化的電腦及儲存系統中，資料係橫跨被組織成儲存階層的多層儲存來容納及佈署。雖然批量資料(包含用於備份的副本)係較佳地以最密集及最便宜的儲存媒體儲存，但資料必須被經常及迅速存取之資料卻放置在儘管最昂貴卻最快速的層之中。資料儲存之最快速且最昂貴的層係電腦系統之揮發性隨機存取記憶體或RAM，其駐存在緊靠微處理器核心附近，並提供最低等待時間及最高頻寬以供資料的隨機存取之用。逐漸密集且更便宜，但速度較慢的層(具有隨機存取之逐漸高的等待時間及更低的頻寬)包含非揮發性固態記憶體或快閃儲存、硬碟驅動器(HDD)，及最後，磁帶驅動器。

【0004】為了要更有效地儲存和處理成長之資料，電腦產業對資料儲存媒體的密度和速度及對電腦的處理能力持續作改進。然而，資料量的增加遠遠超過計算及資料儲存系統之容量和密度中的增進。來自2014年中之資料儲存產業的統計透露出，在過去幾年中所建立和獲得的新資料包含在世界中曾所獲得之資料的大部分。至今，在世界中所建立之資料總量被估計為超過多個皆位元組(皆位元組

係 10^{21} 個位元組)。資料中之大規模的增加在必須可靠地儲存、處理及通訊此資料之資料儲存、計算及通訊系統上寄予大的需求。此激勵用以壓縮資料之無損資料縮減或壓縮技術的增加使用，以致使資料可以用降低的成本儲存，且同樣地，被有效率地處理和傳達。

【0005】各種無損資料縮減或壓縮技術已出現，且已發展多年。該等技術檢驗資料以搜尋資料中之某種形式的冗餘，並利用該冗餘以實施資料足跡之縮減，而無任何資訊的損失。對於看起來要利用資料中之特定形式的冗餘之給定的技術，所達成之資料縮減的程度根據資料中之特定形式的冗餘被多久發現一次。所需要的是，資料縮減技術能靈活地發現和利用資料中之任何可用的冗餘。因為資料源自廣泛種類的來源和環境且以各種形式，所以在處理此多種資料之通用型無損資料縮減的發展及採用中，引起極大的注意。通用型資料縮減技術係除了字母表之外，無需輸入資料的先前知識之技術；因此，其可被普遍地施加至任何及所有的資料，而無需預先知道資料之結構和統計分布特徵。

【0006】可使用以比較不同的資料壓縮技術之實施的長處度量包含在目標資料集上所達成之資料縮減的程度、伴隨壓縮或縮減所達成的效率及伴隨其中資料被解壓縮和檢索以供未來使用之用的效率。效率度量評估解決方案的效能和成本效益。效能度量包含其中新的資料可被消耗和縮減的產出量或攝取速率、縮減資料所需的潛性或時間、

其中資料可被解壓縮和檢索的產出量或速率及解壓縮和檢索資料所需的潛性或時間。成本度量包含諸如微處理器核心或微處理器採用(中央處理單元採用)所需之任何專用硬體組件的成本、專用暫時記憶體的数量和記憶體頻寬，以及由保持資料的儲存之各種層所需的存取數目和頻寬。請注意的是，縮減資料的足跡且同時提供有效率和迅速的壓縮以及解壓縮和檢索，不僅具有降低要儲存和通訊資料的總成本，而且具有有效率地致能資料的隨後處理之益處。

【0007】目前在產業中正在使用的許多通用型資料壓縮技術衍生自 Abraham Lempel及 Jacob Ziv在1977年中所研發出之 Lempel-Ziv壓縮方法，請參閱例如，Jacob Ziv及 Abraham Lempel，“用於序列資料壓縮的通用演算法(A Universal Algorithm for Sequential Data Compression)”，IEEE資訊理論會報，第IT-23冊，第3號，1977年5月。此方法變成用以致能經由網際網路之有效率資料傳輸的基礎。此 Lempel-Ziv方法(命名為LZ77、LZ78及其變異)藉由參照序列呈現之輸入資料流的滑動窗口內所看到的前一事件來置換字串的反覆發生，而縮減資料足跡。在消耗來自輸入資料流之給定資料區塊的新字串時，該等技術透過在目前的及在多達窗口長度之先前的區塊內所預見之所有字串而搜尋。如果新字串係重複的，則可藉由對原始字串之向後參照來置換。若由於重複的字串而被消除之位元組的數目係比用於向後參照所需之位元組的數目更大時，則已獲得資料的縮減。為透過在窗口中所看到的所有字串而搜

尋，且為提供最大的字串匹配，該等技術的實施採用各種方案，包含反覆的掃描，以及建立暫時簿記結構，其包含在窗口中所看到的所有字串之詞典。在消耗輸入的新位元組而組裝新的字串時，該等技術透過在現有窗口中之所有位元組而搜尋，或參照字串之詞典(其次是某種計算)，以決定重覆是否已被發現並以向後參照置換它(或選擇性地，決定是否須對詞典作添加)。

【0008】Lempel-Ziv壓縮方法係經常由施加至資料的第二最佳化所伴隨著，其中源符號係根據其在將被壓縮的資料區塊中之發生的頻率或機率而被動態地重編碼，此常常採用可變寬度編碼方案，以致使較短長度的碼被使用於較頻繁的符號，而藉以導致資料之縮減。例如，請參閱David A. Huffman之“用於最小冗餘碼之建立的方法”，IRE-無線電工程師學會之會議記錄，1952年9月，第1098至1101頁。此技術被稱作Huffman重編碼，且一般需要透過資料用以計算頻率的第一行程和用以實際編碼資料的第二行程。沿此主題的若干變化例亦正在被使用中。

【0009】使用該等技術的範例係已知為“緊縮(Deflate)”之方案，其結合Lempel-Ziv LZ77壓縮方法與Huffman重編碼。緊縮(Deflate)提供壓縮流資料形式規格，其指明用以表示位元組的序列成(通常較短的)位元的序列之方法，以及提供用以包裝後者位元序列成為位元組之方法。該緊縮方案係由PKWARE公司之Phillip W. Katz所原始設計，用於PKZIP歸檔功用。例如，請參閱“字串搜

尋器及使用其之壓縮器”，Phillip W. Katz，美國專利5,051,745，1991年9月24日。美國專利5,051,745描述用以搜尋用於預定目標字串(輸入字串)之符號向量(窗口)的方法。該解決方法對窗口中的各符號採用具有指引符之指引符陣列，以及使用雜湊方法以過濾窗口中之針對輸入字串的相同拷貝而需被搜索的可能位置。此之後係在該等位置處的掃描和字串匹配。

【0010】該緊縮方案係在zlib庫中實施用於資料壓縮。zlib庫係軟體庫，其係諸如Linux、Mac OS X、iOS及各種遊戲機之若干軟體平台的關鍵組件。該zlib庫提供緊縮(Deflate)壓縮及解壓縮碼，以供通過zip(檔案歸檔)、gzip(單一檔案壓縮)、png(用於無損壓縮影像的攜帶式網路圖形)及許多其它應用的使用。目前，zlib係廣泛地使用於資料傳輸和儲存。通過伺服器及瀏覽器的大多數HTTP事務使用zlib以壓縮及解壓縮資料。類似的實施正漸增地由資料儲存系統所使用。

【0011】由英特爾公司(Intel Corp.)在2014年4月中所出版之命名為“在英特爾架構處理器上之高效能ZLIB壓縮”的文章描繪出在當代的英特爾處理器(核心I7 4770處理器，3.4GHz，8MB快取)上運行且在資料的卡爾加里(Calgary)語料庫上操作之zlib最佳版本的壓縮和效能之特徵。在zlib中所使用之緊縮(Deflate)形式設定用於匹配之最小的字串長度為3個字元、該匹配之最長的長度為256個字元及窗口的大小為32k位元組。該實施提供用於9層次之

最佳化的控制，而層次9提供最高的壓縮，但使用最多的計算且執行最徹底的字串匹配，以及層次1係最快的層次並採用貪婪式字串匹配。該文章報告使用zlib層次1(最快層次)之51%的壓縮比，該zlib層次1使用單一線程處理器且耗費每輸入資料位元組17.66個時脈的平均值。在3.4GHz之時脈頻率處，此意味著當用盡單一微處理器核心時之192MB/秒的攝取速率。該報告亦描繪的是，在壓縮中使用小幅上揚之最佳化層次6，效能如何迅速下降到38MB/秒的攝取速率，以及使用最佳化層次9，如何下降到16MB/秒的攝取速率(每位元組209.5個時脈的平均值)。

【0012】現有的資料壓縮解決方法一般使用當代微處理器上之單一處理器核心，而在範圍自10MB/秒至200MB/秒的攝取速率操作。為了要進一步抬高攝取速率，多重核心被使用，或窗口大小被降低。甚至使用定製硬體加速器來達成對該攝取速率的進一步改善，儘管在成本上增加。

【0013】上述之現有資料壓縮方法在利用一般為單一信息或檔案或者幾個檔案的大小之局部窗口中的短字串及符號之程度的細粒度冗餘時係有效的。該等方法在它們被使用於操作在大或極大之資料集上，且需要高速率之資料攝取及資料檢索的應用之中時，具有嚴重的限制和缺點。

【0014】一種重大限制在於，該等方法的實際實施僅可在局部窗口內有效率地利用冗餘。雖然該等方法可接受任意長的輸入資料流，但效率決定的是限制被置放在其中細粒度冗餘將在該窗口的大小範圍上被發現。該等方法係

高度計算密集的，且對窗口中的所有資料需要頻繁及快速的存取。各種簿記結構之字串匹配及查找係在消耗所建立新輸入字串的輸入資料之各個新的位元組(或幾個位元組)時觸發。為了要達成所需的攝取速率，用於字串匹配之窗口及相關聯的機械必須大多駐存在處理器快取子系統中，而實際上，此將對於窗口大小施加約束。

【0015】例如，要在單一處理器核心上達成200MB/秒的攝取速率，每攝取之位元組的平均可用時間預算(包含資料存取和計算)係5奈秒(ns)，其意指使用具有3.4GHz之操作頻率的當代處理器之17個時脈。此預算適合於對晶片上快取，其次，一些字串匹配的存取。目前的處理器具具有數百萬位元組之容量的晶片上快取。對主記憶體之存取耗費超過200個循環(約70奈秒)，所以大部分駐存在記憶體中之較大的窗口將使攝取速率進一步地變慢。此外，因為窗口大小增加，且到重複之字串的距離增加，所以要指明向後參照的成本增加，因而在針對重複的寬廣領域範圍僅只激勵較長的字串將被搜尋。

【0016】在大多數的當代資料儲存系統上，橫跨儲存階層的各種層所存取之資料的足跡係比系統中之記憶體容量更大數個數量級。例如，雖然系統可提供數百個十億位元組的記憶體，但駐存在快閃儲存中之活躍資料的資料足跡可係在數十個萬億位元組之中，以及在儲存系統中的總資料可係在數百個萬億位元組到多個千兆位元組的範圍中。此外，對隨後之儲存層的資料存取之可達到的產出量

會依用於各接續層之數量級或更多而下降。當滑動窗口變得如此之大以致使其不再能配合記憶體時，該等技術將由於顯著低的頻寬，和對下一個資料儲存層次之隨機IO(輸入或輸出操作)存取的較高等待時間而被壓抑。

【0017】例如，考慮到4k位元組之輸入資料的檔案或頁面，其可藉由對，比方說，早已存在於資料中且被散佈在256萬億位元組足跡的範圍之平均長度40個位元組的100個字串做成參照，而由現有資料所組裝。各個參照將花費6個位元組以指明其位址，及1個位元組以供字串長度之用，而有希望節省40個位元組。雖然在此範例中所敘述的頁面可被壓縮超過五倍，但用於此頁面之攝取速率將受到需對儲存系統擷取及驗證100個重複字串之100或更多個IO存取所限制(即使可完全地且廉價地預測到該等字串駐存在何處)。對於僅只10MB/秒的攝取速率，可提供250,000個隨機IO存取/秒(此意指對4KB頁面之1GB/秒的隨機存取頻寬)之儲存系統在用盡該儲存系統之全部頻寬的同時，僅可每秒壓縮2,500個4KB大小的該等頁面，而使其不可用做儲存系統。

【0018】具有萬億位元組或千兆位元組的數量級大小之大窗口之習知壓縮方法的實施，將由於對儲存系統的資料存取之降低的頻寬而感到不足，且將不可接受地變慢。因此，只要冗餘局部地存在於配合處理器快取或系統記憶體的窗口大小上，該等技術的實施就會有效率地發現和利用它。若冗餘係與輸入資料空間地或暫時地分離多個萬億

位元組、千兆位元組或艾(千千兆)位元組時，該等實施將無法以將由儲存存取頻寬所限制之可接受的速率發現該冗餘。

【0019】習知方法之另一限制在於它們並不適用於資料的隨機存取。橫跨整個窗口之所壓縮的資料區塊必須在其中任一區塊內之任一資料塊可被存取之前，被解壓縮。此在窗口的大小上置放實際的限制。此外，其係在未壓縮的資料上傳統地執行之操作(例如，搜尋操作)無法在壓縮的資料上有效率地執行。

【0020】習知方法(且具體地，Lempel-Ziv為主的方法)的又另一限制在於它們僅沿著一維而搜尋冗餘---藉由向後參照而置換相同的字串之該者。Huffman重編碼方案之限制在於其需要透過資料的兩行程，用以計算頻率，並且接著重編碼。這在較大的區塊上會變慢。

【0021】在資料的全域儲存範圍偵測長的重複字串之資料壓縮方法經常使用數位指紋圖譜及雜湊方案的組合。此壓縮處理係稱作重複資料刪除。重複資料刪除之最基本技術將檔案拆散成固定大小的區塊，以及在資料儲存庫的範圍搜索重複的區塊。若檔案的副本被產生時，則在第一檔案中之各區塊將在第二檔案中具有複製品，且該複製品可以用對原始區塊之參照來置換。為了要加速潛在重複區塊的匹配，係採用雜湊之方法。雜湊函數係一種轉換字串成為被稱作其雜湊值的數值之函數。假如兩個字串係一樣的，則它們的雜湊值亦係相等的。雜湊函數映射多個字串

至給定的雜湊值，而可藉以縮減長的字串成極短長度之雜湊值。雜湊值之匹配將比兩個長的字串之匹配更快；因此，該等雜湊值的匹配會先被完成以過濾可係複製品之可能的字串。若輸入之字串或區塊的雜湊值與存在於儲存庫中之字串或區塊的雜湊值匹配時，則可將輸入之字串與在儲存庫中之具有相同雜湊值的各字串相比較，用以確認複製品的存在。

【0022】將檔案拆散成固定大小的區塊係簡單及便利的，且固定大小的區塊係在高效能儲存系統中非常需要的。然而，此技術具有在其可揭露之冗餘數量中的限制，其意指的是，該等技術具有低的壓縮程度。例如，若將第一檔案之副本做成以產生第二檔案時，且若將甚至單一位元組之資料插入至該第二檔案時，則所有下游區塊的對齊將改變，各新的區塊之雜湊值將被重新計算，以及該重複資料刪除方法將不再找出所有的複製品。

【0023】為解決重複資料刪除方法中之此限制，業界已採用指紋圖譜來使匹配內容之位置處的資料流同步及對齊。此後者方案根據指紋圖譜而導致可變大小的區塊。**Michael Rabin**顯示隨機選擇之不可分解多項式可如何被用來採集位元串指紋圖譜---請參閱例如，**Michael O. Rabin**之“藉由隨機多項式之指紋圖譜”，美國哈佛大學計算技術研究中心，TR-15-81，1981年。在此方案中，隨機選擇之質數 p 係使用來藉由計算被視為大的整數模數 p 之該字串的殘留，而採集長字串指紋圖譜。此方案需要在 k 位元整數

上執行整數運算，其中 $k=\log_2(p)$ 。選擇性地，可使用隨機不可分解之 k 次質多項式，以及指紋圖譜係該質多項式之資料模數的多項式表示。

【0024】此指紋圖譜之方法係使用於重複資料刪除系統中，用以識別其中要建立資料塊邊界的合適位置，使得該系統可在全域儲存庫中搜尋該等資料塊的複製品。資料塊邊界可在找到特定值的指紋圖譜時被設定。作為該用法之範例，指紋圖譜可藉由採用32次或更低次的多項式來計算用於輸入資料中之各個及每個48位元組的字串(在輸入的第一個位元組開始，且其次，在之後的每個接續之位元組)。然後，可檢驗32位元指紋圖譜之低的13位元，且無論何時只要該等13位元的值係預定值(例如，值1)，就可設定斷點。對於隨機資料而言，具有該特殊值之該13位元的可能性將係 2^{13} 分之1，以致使該斷點可能要大約每8KB才會被遭遇一次，而導致平均大小8KB的可變大小資料塊。斷點或資料塊邊界將與根據資料之內容的指紋圖譜有效地對齊。當一段長長的距離並沒有發現到指紋圖譜時，則可在某一預定的臨限值處強制斷點，以致使系統一定會建立出比預定大小更短的資料塊，以供儲存庫之用。請參閱例如，Athicha Muthitachareon，Benjie Chen，和David Mazieres之“低頻寬的網路檔案系統”，SOSP'01，第18屆ACM在作業系統原理上之研討會的會議記錄，2001年10月21日，第174至187頁。

【0025】由Michael Rabin及Richard Karp所研發之

Rabin-Karp字串匹配技術對指紋圖譜及字串匹配之效率提供了進一步的改善(請參閱例如，Michael O. Rabin及R. Karp之“有效率的隨機化圖案匹配演算”，IBM Jour. of Res. And Dev.，第31冊，1987年，第249至260頁)。請注意的是，檢驗 m 位元組子字串用於其指紋圖譜之指紋圖譜方法可以用 $O(m)$ 時間評估指紋圖譜多項式函數。因為此方法將需被施加在起始於例如， n 位元組輸入流之每個位元組處的子字串上，所以要在全部資料流上執行指紋圖譜所需之總工作量將係 $O(n \times m)$ 。Rabin-Karp找出被稱作滾動雜湊的雜湊函數，其中可藉由僅做成恆定數目之操作而從前面一者來計算下一個子字串的雜湊值，與該子字串的長度無關。因此，在向右移位一位元組之後，可增量地完成新的 m 位元組字串上之指紋圖譜的計算。此降低了用以計算指紋圖譜的工作量至 $O(1)$ ，及用以採集全部資料流之指紋圖譜的總工作量至 $O(n)$ ，而與資料的大小成線性。此顯著地加速該等指紋圖譜的計算和識別。

【0026】用於上述重複資料刪除方法之典型的資料存取和計算要件可被敘述如下。對於給定之輸入，一旦完成指紋圖譜而建立出資料塊，且在計算出用於該資料塊的雜湊值之後，該等方法首先需要一組對記憶體及隨後儲存層的存取，用以在儲存庫中搜尋及查找保持所有資料塊之雜湊值的全域雜湊表。此將一般需要對儲存之第一IO存取。一旦在雜湊表之中匹配時，則第二組之儲存IO(通常為一個，但可根據儲存庫中存在有多少具有相同雜湊值的資料

塊而超過一個)就隨後擷取載有相同雜湊值的實際資料塊。最後，執行位元組接著位元組之匹配，用以比較輸入資料塊與所擷取的資料塊，而確認及識別出複製品。此係由用於以對原始者的參照置換新的複製品區塊之(對於元資料空間)第三儲存IO存取所跟隨在後。若在全域雜湊表中並無匹配時(或假如並未發現複製品時)，則系統需要一IO，用以登入新的區塊至儲存庫內，以及另一IO，用以更新全域雜湊表而以新的雜湊值登入。因此，對於大的資料集(其中元資料及全域雜湊表不適合在記憶體中，且因此，需要儲存IO以存取它們)，該等系統可需要平均每輸入資料塊三個IO。進一步的改善可藉由採用各種濾波器，使得在全域雜湊表中之錯失可經常被偵測到，而無需用以存取全域雜湊表之第一儲存IO，藉以使處理某些資料塊所需的IO數目減低至兩個。

【0027】可提供250,000個隨機IO存取/秒(此意指對4KB頁面之1GB/秒的隨機存取頻寬)之儲存系統在用盡該儲存系統之全部頻寬的同時，可攝取和重複資料刪除大約每秒83,333個(250,000除以每輸入資料塊三個IO)平均大小4KB的輸入資料塊，而致能333MB/秒的攝取速率。若僅使用儲存系統的一半頻寬時(使得另一半頻寬係可用於對所儲存資料之存取)，該重複資料刪除系統仍可傳遞166MB/秒的攝取速率。假定足夠的處理功率係可用於系統之中時，則該等攝取速率(其係由I/O頻寬所限制)係可達成的。因此，給定足夠的處理功率，重複資料刪除系統能在當代

儲存系統上，以經濟的IO在資料之全域範疇查找大的資料複製品，且以每秒數百個百萬位元組之攝取速率傳遞資料縮減。

【0028】根據上述說明，應瞭解的是，雖然該等重複資料刪除方法係在查找全域範圍之長字串的複製品時有效，但它們主要在查找大的複製品時有效。若資料具有更細粒度的變化或修正時，則將不會使用此方法來查找可用的冗餘。此大大降低了該等方法有效的範圍之資料集的寬度。該等方法已在某些資料儲存系統和應用中發現用途，例如，資料的正規備份，其中將予以備份之新資料僅有一些檔案被修正，且其餘的部分均係在之前的備份中所儲歸檔案之複製品。同樣地，重複資料刪除為主的系統常常被部署在其中多個精確拷貝之資料或碼被做成的環境中，諸如在資料中心之中的虛擬化環境中。然而，當資料更普遍地或更細粒度地演變和被修正時，則重複資料刪除為主的技術將喪失其效用。

【0029】某些(通常在資料備份應用中被採用的)方法並不執行輸入資料與其雜湊值匹配該輸入的該者之間的實際位元組接著位元組之比較。該等解決方法取決使用如同SHA-1之強大雜湊函數之碰撞的低機率。然而，由於碰撞之有限的非零機率(其中多個不同的字串可映射至相同的雜湊值)，所以該等方法不能被視為提供無損資料縮減，且因此，將不符合最初儲存及通訊的高資料完整性需求。

【0030】某些方法結合多個現有的資料壓縮技術。通

常，在該安裝中，全域重複資料刪除方法首先施加至資料。之後，在該重複資料刪除的資料集上，及採用小的窗口，與 Huffman 重編碼結合之 Lempel-Ziv 字串壓縮方法被施加以達成進一步的資料縮減。

【0031】然而，儘管採用所有迄今已知之技術，在成長及累積資料的需求，與世界經濟可使用最佳可用的現代儲存系統實惠地容納甚麼之間，仍持續有若干數量級的差距。給定由該等成長資料所需求之儲存容量的特殊要求，對於用以進一步縮減資料足跡之改善方式，就持續有所需要。因此，仍需要發展出解決現有技術之限制，或沿著尚未由現有技術所解決的層面以利用資料中之可用冗餘的方法。同時，能以可接受的處理速率和處理成本有效率地存取及檢索資料就持續變得重要。還需要能夠直接對縮減的資料進行有效的搜索操作。

【0032】綜上所述，對於可利用在大和極大資料集範圍之冗餘，及可提供高速率之資料攝取和資料檢索的無損資料縮減解決方法，仍有長期的需求。

【發明內容】

【0033】此處所敘述之實施例以技術及系統為特徵，其可對於大和極大之資料集執行無損資料縮減，且同時提供高速率的資料攝取及資料檢索，以及其並不會因現有的資料壓縮系統的缺點及限制而受損。

【0034】具體地，一些實施例可以將來自視頻資料的

壓縮動畫資料和壓縮音頻資料解壓縮。接著，實施例可以將來自該壓縮動畫資料的內訊框(I訊框)解壓縮。接著，實施例可以無損縮減該I訊框以獲得無損縮減的I訊框。對於每個I訊框，該無損縮減的I訊框可以包含(1)藉由使用該I訊框來識別第一組主要資料元件，以對於根據其內容組織主要資料元件的資料結構執行第一內容關聯查找，以及(2)使用該第一組主要資料元件來無損縮減該I訊框。實施例可以額外地將壓縮的音頻資料解壓縮，以獲得一組音頻組件。接下來，對於該組音頻組件中的每一個音頻組件，實施例可以(1)藉由使用該音頻組件來識別第二組主要資料元件，以對於根據其內容組織主要資料元件的該資料結構執行第二內容關聯查找，以及(2)使用該第二組主要資料元件來無損縮減該音頻組件。

【0035】一些實施例可以將儲存在第一記憶體裝置中且被配置成根據其內容來組織主要資料元件的資料結構初始化。接下來，實施例可以將輸入資料分解成一連串的候選元件。對於每個候選元件，實施例可以(1)藉由使用該候選元件來對於該資料結構執行內容關聯查找以識別一組主要資料元件，以及(2)藉由使用該組主要資料元件來無損縮減該候選元件，其中如果該候選元件的大小沒有被充分縮減，則將該候選元件加入到該資料結構作為新的主要資料元件。接下來，實施例可以將該無損縮減的候選元件儲存在第二記憶體裝置中。在檢測到該資料結構的一或多個組件的大小係大於臨限值時，實施例可以(1)將該資料

結構的一或多個組件移動到該第二記憶體裝置，以及(2)將被移動到該第二記憶體裝置之該資料結構的該一或多個組件初始化。無損縮減的資料批量可以包含(1)在時間上相鄰的初始化之間被儲存在該第二記憶體裝置上的無損縮減的候選元件，以及(2)在該時間上相鄰的初始化之間被移動到該第二記憶體裝置的該資料結構的組件。在變型中，實施例可以根據儲存在該第二記憶體裝置上的無損縮減的資料批量建立一組包裹，其中該組包裹有利於資料從一台電腦到另一台電腦的歸檔和移動。

【0036】一些實施例可以將輸入資料分解成一連串的候選元件。接下來，對於每個候選元件，實施例可以(1)將該候選元件拆分成一或多個欄位，(2)對於每個欄位，將該欄位除以質多項式，以獲得商和餘數對，(3)根據一或多個商和餘數對來確定名稱，(4)藉由使用該名稱來識別一組主要資料元件，以對於根據其各自名稱的內容來組織主要資料元件的資料結構執行內容關聯查找，以及(5)藉由使用該組主要資料元件來無損縮減該候選元件。

【0037】一些實施例可以將輸入資料分解成一連串的候選元件。接下來，對於每個候選元件，實施例可以(1)藉由使用該候選元件來識別一組主要資料元件，以對於根據其內容來組織主要資料元件的資料結構執行內容關聯查找，以及(2)藉由使用該組主要資料元件來無損縮減該候選元件。接著，實施例可以將無損縮減的候選元件儲存在一組提取檔案中。接下來，實施例可以將該主要資料元件

儲存在一組主要資料元件檔案中。在一些實施例中，每個無損縮減的候選元件針對用於縮減該候選元件的每個主要資料元件指定含有該主要資料元件的主要資料元件檔案和可以在該主要資料元件檔案中找到該主要資料元件的偏移量。在一些實施例中，每個提取檔案儲存含有主要資料元件的主要資料元件檔案的列表，該主要資料元件用於無損縮減儲存在該提取檔案中的候選元件。

【0038】 使用該組主要資料元件來無損縮減資料元件(例如，I訊框、音頻組件、候選元件等)可包含：(1)響應於確定(i)對於該組主要資料元件之參照的大小和(ii)重建程式的描述的大小之總和係小於該候選元件的大小的臨限值部分，產生該資料元件的第一無損縮減表示，其中該第一無損縮減表示包含對於在該組主要資料元件中的每個主要資料元件的參照和該重建程式的描述；(2)響應於確定(i)對於該組主要資料元件之該參照的該大小和(ii)該重建程式該描述的該大小之該總和係大於或等於該資料元件的該大小的該臨限值部分，在該資料結構中加入該資料元件作為新的主要資料元件，以及產生該資料元件的第二無損縮減表示，其中該第二無損縮減表示包含對於該新的主要資料元件的參照。注意，該重建程式的該描述可以指定轉換的序列，其中當施加到該組主要資料元件時(即，用於無損地減少資料元件的一或多個主要資料元件)，產生該資料元件。

【圖式簡單說明】

【0039】 第1A圖顯示根據在此所敘述的一些實施例之用於資料縮減的方法及設備，其將輸入資料分解成元件，及自駐存在主要資料節中之主要資料元件得到該等者。

【0040】 第1B至1G圖顯示根據在此所敘述的一些實施例之被描繪在第1A圖中的方法及設備之變化例。

【0041】 第1H圖呈現根據在此所敘述的一些實施例之形式及規格的範例，其描繪提取之資料的結構。

【0042】 第1I至1P圖顯示輸入資料成無損縮減形式之概念上的轉變，用於第1A至1G圖中所示之用於資料縮減的方法及設備之變化例。

【0043】 第2圖顯示根據在此所敘述的一些實施例之藉由將輸入資料分解成元件，及自駐存在主要資料節中之主要資料元件取得該等元件之用於資料縮減的處理。

【0044】 第3A、3B、3C、3D及3E圖顯示根據在此所敘述的一些實施例之不同的資料組織系統，其可使用以根據主要資料元件名稱來組織它們。

【0045】 第3F圖呈現根據在此所敘述的一些實施例之自描述樹狀節點資料結構。

【0046】 第3G圖呈現根據在此所敘述的一些實施例之自描述葉狀節點資料結構。

【0047】 第3H圖呈現根據在此所敘述的一些實施例之自描述葉狀節點資料結構，其包含導航預看欄位。

【0048】 第4圖顯示根據在此所敘述的一些實施例之

256TB的主要資料可如何以樹狀形式組織之範例，並呈現該樹狀物可如何被佈局在記憶體及儲存器中。

【0049】第5A至5C圖顯示資料可如何使用在此所敘述之實施例而被組織的實際範例。

【0050】第6A至6C圖顯示根據在此所敘述的一些實施例之樹狀資料結構可如何使用於參照第1A至1C圖所分別敘述的內容相關映射器。

【0051】第7A圖提供根據在此所敘述的一些實施例之可在重建程式中被指明的轉變之範例。

【0052】第7B圖顯示根據在此所敘述的一些實施例之將從主要資料元件得到的候選元件之結果的範例。

【0053】第8A至8E圖顯示根據在此所敘述的一些實施例之資料縮減可如何藉由將輸入資料分解成固定大小元件，且以參照第3D及3E圖所敘述的樹狀資料結構組織該等元件來執行。

【0054】第9A至9C圖顯示根據在此所敘述的一些實施例之基於第1C圖中所示系統的資料提取(Data Distillation™)方案之範例。

【0055】第10A圖提供根據在此所敘述的一些實施例之在重建程式中所指明的轉變如何施加至主要資料元件以產生衍生物元件之範例。

【0056】第10B至10C圖顯示根據在此所敘述的一些實施例之資料檢索處理。

【0057】第11A至11G圖顯示根據在此所敘述的一些

實施例之包含資料提取(Data Distillation™)機制(其可使用軟體、硬體或它們的組合來實施)的系統。

【0058】第11H圖顯示根據在此所敘述的一些實施例之資料提取(Data Distillation™)設備可如何與取樣通用型計算平台介接。

【0059】第11I圖顯示如何將資料提取(Data Distillation™)設備用於在區塊處理儲存系統中的資料縮減。

【0060】第12A至12B圖顯示根據在此所敘述的一些實施例之用於橫跨頻寬約束通訊媒體的資料通訊之資料提取(Data Distillation™)設備的使用。

【0061】第12C至12K圖顯示根據在此所敘述的一些實施例之藉由用於各種使用模型的資料提取(Data Distillation™)設備而產生之縮減資料的各種組件。

【0062】第12L至R圖根據在此描述的一些實施例顯示提取處理可以在分散式系統上如何部署和執行，以便能夠容納在非常高的攝取率的非常大的資料集。

【0063】第13至17圖根據在此描述的一些實施例顯示多維度搜索和資料檢索可對縮減資料如何進行。

【0064】第18A至18B圖顯示用於根據MPEG 1、Layer 3標準(也稱為MP3)的音頻資料壓縮和解壓縮的編碼器和解碼器的方塊圖。

【0065】第18C圖顯示在第1A圖首先顯示的資料提取設備如何可以被強化以對於MP3資料執行資料縮減。

【0066】第19圖顯示在第1A圖首先顯示的資料提取設備如何可以被強化以對於視頻資料執行資料縮減。

【實施方式】

【0067】以下說明係提出用以使熟習本領域之任何人士能做成及使用本發明，且係以特定應用及其需求之情境來提供。對所揭示之實施例的各種修正將對於本領域技術人員是顯而易見的，且在此所界定的一般原理可被應用至其它的實施例和應用，而不會背離本發明之精神及範疇。因此，本發明並未受限於所示之該等實施例，且係符合與在此所揭示之原理及特性一致的更寬廣範圍。在此發明中，當詞組使用具有一組實體之的用語“及／或”時，該詞組涵蓋該組實體的所有可能之組合，除非另有指明。例如，詞組“X，Y，及／或Z”將涵蓋以下七種組合：“僅X”、“僅Y”、“僅Z”、“X及Y，但非Z”、“X及Z，但非Y”、“Y及Z，但非X”、以及“X，Y，及Z”。

使用主要資料節之有效率的資料無損縮減

【0068】在此處所敘述之一些實施例中，資料被組織及儲存，用以在整個資料集的範圍有效率地全域揭露和利用冗餘。輸入資料流係拆散成被稱作元件的構成片或資料塊，且在元件之中的冗餘係以比元件本身更細的粒狀偵測及利用，藉以降低所儲存之資料的總體足跡。稱作主要資料元件的一組元件係識別且使用為用於資料集之共同及共

享的構建模組，以及儲存在被稱作主要資料篩或主要資料儲存的結構中。主要資料元件單純是某一大小之位元、位元組或數字的序列。主要資料元件可根據實施而成為固定大小或可變大小的。輸入資料的其它構成元件係源自主要資料元件，且被稱做衍生物元件。因此，輸入資料被分解成主要資料元件及衍生物元件。

【0069】 主要資料篩佈置及組織主要資料元件，以致使主要資料篩可被以內容相關方式搜尋和存取。給定具有一些限制之一些輸入內容，主要資料篩可被查詢以檢索具有該內容的主要資料元件。給定輸入元件，主要資料篩可使用該元件的值，或該元件中之某些欄位的值，以快速提供一或小的組之主要資料元件，其中輸入元件可以用用以指明衍生所需的最小儲存自它取得。在一些實施例中，於主要資料篩中之該等元件係以三種形式組織。衍生物元件係藉由對主要資料元件執行轉變而自它取得，該轉變係指明於重建程式中，其描述如何自一或多個主要資料元件產生衍生物元件。距離臨限值指明衍生物元件的儲存足跡之大小上的限制。此臨限值指明衍生物元件與主要資料元件之最大的可允許距離，且亦可對於被用來產生衍生物元件之重建程式的大小給予限制。

【0070】 衍生資料的檢索可藉由對於由衍生所指明之一或多個主要資料元件執行重建程式來完成。

【0071】 在此發明中，上述通用型無損資料縮減技術可被稱作資料提取(Data Distillation™)處理。其執行與化

學中之提取相似的功能---將混合物分離成其構成元件。該主要資料篩亦被稱作篩或資料提取(Data Distillation™)篩。

【0072】在此方案中，輸入資料流被分解成元件之序列，各元件係主要資料元件或衍生自一或多個主要資料元件的衍生物元件。各元件被轉變成無損縮減表示，該無損縮減表示在主要資料元件之情況中包含對該主要資料元件的參照，及在衍生物元件之情況中包含對衍生中所包含之一或多個主要資料元件的參照，並且包含重建程式的說明。因此，輸入資料流係分解成其係在無損縮減表示之中的元件之序列。此元件之序列(出現在無損縮減表示之中)被稱作提取資料流或提取資料。在提取資料中的元件之序列具有相對於輸入資料中的元件之序列的一對一對應關係亦即，在提取資料中的元件之序列中的第n個元件對應輸入資料中的元件之序列中的第n個元件。

【0073】在此發明中所敘述之通用型無損資料縮減技術接收輸入資料流且將其轉換成提取資料流及主要資料篩的組合，以致使提取資料流及主要資料篩之足跡的總和通常比輸入資料流的足跡更小。在此發明中，提取資料流及主要資料篩係統稱為無損縮減資料，且亦將可互換地稱作“縮減資料流”或“縮減資料”或“縮減之資料”。同樣地，對於由在此發明中所敘述之無損資料縮減技術所產生，且以無損縮減之形式出現的元件之序列，以下的用語係可互換地使用：“縮減輸出資料流”、“縮減輸出資料”、“提取資

料流”、“提取資料”及“提取之資料”。

【0074】第1A圖顯示根據在此所敘述的一些實施例之用於資料縮減的方法及設備，其將輸入資料分解成元件，及自駐存在主要資料篩中之主要資料元件取得該等者。此圖式描繪資料縮減或資料提取(Data Distillation™)方法及設備的整體方塊圖，且提供功能性組件、結構及操作的綜覽。在第1A圖中所描繪的組件及／或操作可使用軟體、硬體或其組合來實施。

【0075】位元組之序列係接收自輸入資料流，且呈現為至資料縮減設備103(亦稱作資料提取(Data Distillation™)設備)的輸入資料102。解析器及分解器104解析該輸入資料且將其拆散成候選元件。該分解器決定輸入流之何處應插入分隔符以將該流切片成候選元件。一旦已識別出資料中之兩個接連的分隔符，候選元件105就由解析器及分解器所產生，且被提交至主要資料篩106(亦稱作資料提取(Data Distillation™)篩)。

【0076】資料提取(Data Distillation™)篩或主要資料篩106包含所有的主要資料元件(在第1A圖中被標記為PDE)，並根據它們的值或內容來佈置和組織它們。該篩提供用於兩種存取之支援。第一，主要資料元件的各者可經由對主要資料元件駐存在該篩中之位置的參照來被直接存取。第二，元件可藉由使用可以用軟體、硬體或其組合實施的內容相關映射器121來以內容相關方式存取。此種對篩之第二形式的存取係重要的特性，該特性係由本發明實

施例所使用，用以識別與候選元件 105 確實匹配的主要資料元件，或用以識別其中候選元件可自其取得的主要資料元件。具體地，給定候選元件(例如，候選元件 105)，主要資料篩 106 可被搜尋(根據候選元件 105 的值，或根據在候選元件 105 中之某些欄位的值)，用以從該候選元件可以用用以指明衍生所需的最小儲存而衍生之處來快速提供一或小的組之主要資料元件 107。

【0077】該篩或主要資料篩 106 可以用其值係散佈在資料空間之範圍的一組主要資料元件來初始化。選擇性地，根據在此參照第 1A 至 1C 圖及第 2 圖所敘述之資料提取(Data Distillation™)處理，篩可清空地開始，且當攝取資料時，主要資料元件可動態地被添加至篩。

【0078】衍生器 110 接收候選元件 105 及適合於衍生之檢索的主要資料元件 107(其係內容相關地檢索自主要資料篩 106)，決定候選元件 105 是否可衍生自該等主要資料元件的一或多者，產生縮減資料組件 115(包含對相關聯的主要資料元件和重建程式之參照)，以及提供更新 114 至主要資料篩。若候選元件係檢索的主要資料元件之複製品時，則衍生器置放對位在主要資料篩中之主要資料元件的參照(或指引符)，及此係主要資料元件的指示符至提取之資料 108 內。若並未發現複製品時，則衍生器表示候選元件為被執行於一或多個檢索的主要資料元件上之一或多個轉變的結果，其中轉變之序列係統稱為例如，重建程式 119A 的重建程式。各衍生可要求將由衍生器所建構之其自身獨特

的程式。重建程式指明諸如，可被施加至主要資料元件之插入、刪除、置換、串聯、算術及邏輯操作的轉變。倘若衍生物元件的足跡(被計算為重建程式之大小加上對所需要之主要資料元件的參照之大小)係在相對於候選元件之某一特定的距離臨限值之內時(為致能資料縮減)，則候選元件被重制定成衍生物元件，且由重建程式及對相關聯的主要資料元件(或元件)之參照的組合所置換---在此情況中，該等者形成縮減資料組件 115。若超出該臨限值時，或假如並無合適的主要資料元件從主要資料篩被檢索出時，則主要資料篩可被指示以安裝該候選者成為新的主要資料元件。在此情況中，衍生器將對新添加之主要資料元件的參照，及此係主要資料元件的指示符放置到提取之資料內。

【0079】用於資料之檢索的請求(例如，檢索請求 109)可以用對包含主要資料元件的主要資料篩中之位置的參照，或在衍生物的情況中，對主要資料元件的該參照及相關聯的重建程式(或在根據多個主要資料元件之衍生物的情況中，對該多個主要資料元件的該等參照及相關聯的重建程式)之組合的形式。使用對主要資料篩中之主要資料元件的一或多個參照，檢索器 111 可擷取主要資料篩以提取一或多個主要資料元件，且提供該一或多個主要資料元件以及重建程式至重建器 112，以回應資料檢索請求，該重建器 112 在該一或多個主要資料元件上執行轉變(被指明於該重建程式中)，用以產生重建之資料 116(其係所請

求之資料)並將其傳遞至檢索資料輸出113。

【0080】在此實施例的變化例中，該等主要資料元件可以用壓縮形式被儲存於篩中(使用現有技術中已知之技術，包含Huffman編碼及Lempel Ziv法)，且當需要時，予以解壓縮。此具有縮減主要資料篩之整體足跡的優點。唯一的約束在於，內容相關映射器121必須如之前一樣地提供內容相關存取至該等主要資料元件。

【0081】第1B及1C圖顯示根據在此所敘述的一些實施例之被描繪在第1A圖中的方法及設備之變化例。在第1B圖中，重建程式可予以儲存在主要資料篩中，且可被與主要資料元件相似地處理。對重建程式的參照或指引符119B係設置在提取之資料108中，以取代提供重建程式119A之本身。若該重建程式係由其它衍生物所共享時，且若對重建程式的參照或指引符(加上用以在重建程式與對重建程式的參照之間區分所需的任何元資料)要求比該重建程式之本身更少的儲存空間時，則可獲得進一步的資料縮減。

【0082】在第1B圖中，重建程式可就像主要資料元件一樣地被處理和存取，且儲存在主要資料篩中作為主要資料元件，而藉以允許從主要資料篩之重建程式的內容相關搜尋和檢索。在該衍生處理以建立衍生物元件之期間，一旦衍生器110決定了用於該衍生所需之重建程式，其就可接著決定此候選重建程式是否係早已存在於主要資料篩中，或此候選重建程式是否可自早已存在於主要資料篩中

的另一條目取得。若該候選重建程式係早已存在於主要資料篩之中時，則衍生器 110 可決定對該預先存在之條目的參照，且可將該參照包含在提取之資料 108 中。若候選重建程式可自早已駐存在主要資料篩之中的現有條目取得時，則衍生器可傳遞該候選重建程式的衍生物或重制訂至提取之資料，亦即，衍生器置放對預先存在於主要資料篩中之條目的參照，伴隨增量的重建程式至提取之資料內，該增量的重建程式係從預先存在之條目取得。若該候選重建程式並未存在於主要資料篩之中，也無法自主要資料篩中的條目取得時，則衍生器 110 可添加該重建程式至主要資料篩(添加重建程式至儲存的操作可回報對新添加之條目的參照)，且可將對該重建程式的參照包含在提取之資料 108 中。

【0083】第 1C 圖呈現根據在此所敘述的一些實施例之被描繪在第 1B 圖中的方法及設備之變化例。具體地，被用來儲存及查詢重建程式之第 1C 圖中的機制係與被用來儲存及查詢主要資料元件的機制相似，但重建程式係保持在與包含主要資料元件之結構分離的結構(稱作主要重建程式篩)中。在該結構中之條目係稱作主要重建程式(在第 1C 圖中被標記為 PRP)。回顧一下，主要資料篩 106 包含內容相關映射器 121，其支援快速的內容相關查找操作。在第 1C 圖中所描繪之實施例包含內容相關映射器 122，其係與內容相關映射器 121 相似。在第 1C 圖中，內容相關映射器 122 及內容相關映射器 121 已被顯示成為主要資料篩或主要資

料儲存106的一部分。在其它實施例中，內容相關映射器122及重建程式在稱作主要重建程式篩的結構中可被與主要資料篩或主要資料儲存106分離地儲存。

【0084】在此實施例的變化例中，該等主要資料元件可以用壓縮形式被儲存於篩中(使用現有技術中已知之技術，包含Huffman編碼及Lempel Ziv法)，且當需要時，予以解壓縮。同樣地，主要重建程式可以用壓縮形式被儲存於主要重建程式篩中(使用現有技術中已知之技術，包含Huffman編碼及Lempel Ziv法)，且當需要時，予以解壓縮。此具有縮減主要資料篩及主要重建程式篩之整體足跡的優點。唯一的約束在於，內容相關映射器121及122必須如之前一樣地提供內容相關存取至該等主要資料元件及主要重建程式。

【0085】第1D圖呈現根據在此所敘述的一些實施例之被描繪在第1A圖中的方法及設備之變化例。具體地，在第1D圖中所敘述的實施例中，主要資料元件被內嵌地儲存在提取資料中。主要資料篩或主要資料儲存106持續提供內容相關存取至主要資料元件，且持續邏輯地包含主要資料元件。其保持對提取資料中所內嵌設置的主要資料元件之參照或鏈路。例如，在第1D圖中，主要資料元件130係內嵌地設置在提取之資料108中。主要資料篩或主要資料儲存106保持對主要資料元件130的參照131。再次地，在此設置中，衍生物元件的無損縮減表示將包含對所需之主要資料元件的參照。在資料檢索期間，檢索器111將自所需

之主要資料元件所設置之處擷取該所需之主要資料元件。

【0086】第1E圖呈現根據在此所敘述的一些實施例之被描繪在第1D圖中的方法及設備之變化例。具體地，在第1E圖之中所敘述的實施例中，就像在第1B圖之中所描繪的設置中，重建程式可自主要重建程式取得，且被指明為增量之重建程式加上對主要重建程式的參照。該等主要重建程式有如主要資料元件一樣地被處理，且被邏輯地安裝在主要資料篩之中。再者，在此設置中，主要資料元件和主要重建程式二者係內嵌地儲存在提取資料中。主要資料篩或主要資料儲存106持續提供內容相關存取至主要資料元件及主要重建程式，且持續邏輯地包含主要資料元件及主要重建程式，且同時保持對於它們在提取資料中所內嵌設置之處的參照或鏈路。例如，在第1E圖中，主要資料元件130係內嵌地設置在提取之資料108中。此外，在第1E圖中，主要重建程式132係內嵌地設置在提取資料中。主要資料篩或主要資料儲存106保持對主要資料元件130(其係PDE_i)的參照131(其係對PDE_i的參照)及對主要重建程式132(其係主要重建程式₁)的參照133(其係對PDE_j的參照)。再次地，在此設置中，衍生物元件的無損縮減表示將包含對所需之主要資料元件及所需之主要重建程式的參照。在資料檢索期間，檢索器111將自它們在對應的提取資料中所設置之處擷取該等所需之組件。

【0087】第1F圖呈現根據在此所敘述的一些實施例之被描繪在第1E圖中的方法及設備之變化例。具體地，在第

1F圖之中所敘述的實施例中，就像在第1C圖之中所描繪的設置中，主要資料篩108包含分離的映射器---用於主要資料元件的內容相關映射器121及用於主要重建程式的內容相關映射器122。

【0088】第1G圖呈現被描繪在第1A圖至第1F圖中的方法及設備之更一般性的變化例。具體地，在第1G圖之中所敘述的實施例中，可將主要資料元件設置在主要資料篩之中，或內嵌地設置在提取資料中。可將一些主要資料元件設置在主要資料篩之中，且同時將其它者內嵌地設置在提取資料中。同樣地，可將主要重建程式設置在主要資料篩之中，或內嵌地設置在提取資料中。可將一些主要重建程式設置在主要資料篩之中，且同時將其它者內嵌地設置在提取資料中。主要資料篩邏輯地包含所有的主要資料元件和主要重建程式，且在主要資料元件或主要重建程式係內嵌地設置在提取資料中的情況下，該主要資料篩提供對其位置的參照。

【0089】用於資料縮減之將輸入資料分解成元件，及自駐存在主要資料篩中的主要資料元件取得該等者之方法和設備的上述說明，僅被呈現以供描繪及說明的目的之用。它們並不打算要窮盡或要限制本發明於所揭示的形式。因而，許多修正和變化將顯而易見於熟習本領域之從業者。

【0090】第1H圖呈現根據在此所敘述的一些實施例之形式及規格的範例，其描繪用於資料提取處理的方法和設

備之第1A至G圖中的提取之資料119A的結構。因為資料提取處理將輸入資料分解成主要資料元件及衍生物元件，所以用於資料之無損縮減表示的形式識別該等元件且描述提取資料中的該等元件之各種組件。該自描述形式識別提取資料中的各元件，指示其是否係主要資料元件或衍生物元件，以及描繪該等各種組件，亦即，對被安裝在篩中之一或多個主要資料元件的參照，對被安裝在主要資料篩中之重建程式的參照(如在第1B圖之119B中)或對被安裝在主要重建程式(PRP)篩中之重建程式的參照(如在第1C圖之119C中)，及內嵌的重建程式(RP)。該主要重建程式(PRP)篩亦可被互換地稱作主要重建程式(PRP)儲存。在第1H圖中之形式具有用以藉由對於多個主要資料元件執行重建程式，而以衍生物元件及可獨立指明之主要資料元件之各者的大小來指明衍生之規定。在第1H圖中之形式亦具有用以指明主要資料元件係內嵌地設置在提取資料中，而非被設置在主要資料篩內之規定。此係由操作碼(Opcode)編碼7所指明，其指明的是，元件的類型係主要資料元件，其係內嵌地設置在提取資料中。該提取之資料係使用此形式而被儲存於資料儲存系統中。在此形式中之資料係由資料檢索器111所消耗，以致使資料的各種組件可被擷取且隨後地重建。

【0091】第1I至1P圖顯示將輸入資料概念上轉變成無損縮減形式，用於第1A至1G圖中所示之用於資料縮減的方法及設備之變化例。第1I圖顯示輸入資料之流如何被分

解成候選元件，且隨後候選元件如何被推定為主要資料元件或衍生物元件。最後，資料被轉變成無損縮減形式。第1I至1N圖顯示用於各種實施例之無損縮減形式的變化例。

【0092】第1I圖及第1J圖顯示由第1A圖中所描繪之方法及設備所產生的資料之無損縮減形式的範例。在第1I圖中的無損縮減形式包含內容相關映射器，且係致能對現有之主要資料元件的資料之連續進一步攝取及此資料之縮減的形式；另一方面，在第1J圖中的無損縮減形式不再保留內容相關映射器，而導致較小的資料足跡。第1K圖及第1L圖顯示由第1C圖中所描繪之方法及設備所產生的資料之無損縮減形式的範例。在第1K圖中的無損縮減形式包含內容相關映射器，且係致能對現有之主要資料元件及主要重建程式的資料之連續進一步攝取及此資料之縮減的形式；另一方面，在第1L圖中的無損縮減形式不再保留內容相關映射器，而導致較小的資料足跡。

【0093】第1M圖至第1N圖顯示由第1F圖中所描繪之方法及設備所產生的資料之無損縮減形式的範例，其中主要資料元件及主要重建程式係內嵌地設置在提取資料中。在第1M圖中的無損縮減形式包含內容相關映射器，且係致能對現有之主要資料元件及主要重建程式的資料之連續進一步攝取及此資料之縮減的形式；另一方面，在第1N圖中的無損縮減形式不再保留內容相關映射器，而導致較小的資料足跡。第1O圖及第1P圖顯示由第1G圖中所描繪之方法及設備所產生的資料之無損縮減形式的範例，其中主

要資料元件及主要重建程式係內嵌地設置在提取資料中或在主要資料篩中。在第1O圖中的無損縮減形式包含內容相關映射器，且係致能對現有之主要資料元件及主要重建程式的資料之連續進一步攝取及此資料之縮減的形式；另一方面，在第1P圖中的無損縮減形式不再保留內容相關映射器，而導致較小的資料足跡。

【0094】在第1A至1P圖中所示之實施例的變化例中，縮減之資料的各種組件可使用先前技術中所已知之技術(諸如Huffman編碼，及Lempel Ziv法)來進一步縮減或壓縮，且以此壓縮形式儲存。該等組件可當它們需用於資料提取設備中之使用時，被隨後地解壓縮。此具有進一步縮減資料之整體足跡的優點。

【0095】第2圖顯示根據在此所敘述的一些實施例之藉由將輸入資料分解成元件，及自駐存在主要資料篩中之主要資料元件取得該等元件之用於資料縮減的處理。當輸入資料到達時，其可被解析且分解或拆散成一串列的候選元件(操作202)。接著，候選元件係消耗自輸入(操作204)，以及主要資料篩的內容相關查找係根據候選元件的內容來執行，用以查明其中可取得該候選元件之任何合適的元件是否存在(操作206)。若主要資料篩並未發現任何該等元件時(操作208的“否”分支)，則該候選元件將被配置及進入到篩內，作為新的主要資料元件，且在提取資料中被建立用於該候選元件的條目將變成對該新建立之主要資料元件的參照(操作216)。若主要資料篩的內容相關查找

產生可潛在地取得該候選元件之一或多個合適的元件時(操作208的“是”分支)，則在所檢索出之主要資料元件上執行分析及計算，以從它們取得該候選元件。應注意的是，在一些實施例中，僅用於合適之主要資料元件的元資料被首先擷取，且對於該元資料執行分析，而如果被推定有用，合適的主要資料元件才會被隨後擷取(在該等實施例中，用於主要資料元件的元資料提供有關該主要資料元件之內容的一些資訊，而藉以使系統根據該元資料而快速排除匹配或評估可取得性)。在其它實施例中，主要資料篩直接檢索主要資料元件(亦即，無需在檢索主要資料元件之前，先檢索元資料而分析該元資料)，所以對於所檢索出的主要資料元件執行分析及計算。

【0096】執行第一檢查，以查明該候選者是否係該等元件之任一者的複製品(操作210)。此檢查可使用任一合適的雜湊技術來加快。若該候選者係與檢索自主要資料篩之主要資料元件相同時(操作210的“是”分支)，則在提取資料中被建立用於該候選者的條目係由對此主要資料元件的參照，及此條目係主要資料元件的指示所置換(操作220)。若並未發現到複製品時(操作210的“否”分支)，則根據該候選元件而檢索自主要資料篩之該等條目被視為其中可潛在地取得該候選元件的條目。以下係主要資料篩之重要的、新穎的及非顯而易見的特徵：當複製品並未在主要資料篩之中被發現時，該主要資料篩可回報的是，雖然與候選元件不相同，但主要資料元件係候選元件可藉由對於

該主要資料元件施加一或多個轉變而被潛在地取得的元件。該處理可接著執行分析及計算，用以自最合適的主要資料元件或一組合適的主要資料元件取得候選元件(操作 212)。在一些實施例中，衍生將候選元件表示為對於一或多個主要資料元件執行轉變的結果，該等轉變係統稱為重建程式。各衍生可要求其自身獨特的程式被建構。除了建構重建程式之外，該處理亦可計算距離度量，其一般指示用以儲存候選元件的重制定及自該重制定重建該候選元件所需之儲存資源及/或計算資源的層次。在一些實施例中，衍生物元件的足跡係使用作為候選者與主要資料元件之距離的測度---具體地，距離度量可被界定為重建程式之大小加上對該衍生中所涉及之一或多個主要資料元件的參照之大小的總和。具有最短距離的衍生可被選擇。用於此衍生之距離係與距離臨限值相比(操作 214)，且若該距離並未超過距離臨限值時，則該衍生被接受(操作 214的“是”分支)。為了要產生資料縮減，距離臨限值必須一直小於候選元件的大小。例如，距離臨限值可被設定為候選元件之大小的 50%，使得只在衍生的足跡係小於或等於候選元件的足跡的一半時，該衍生才被接受，而藉以確保合適的衍生存在之用於各候選元件之兩倍或更大的縮減。該距離臨限值可係根據使用者特定輸入或由系統所選擇的預定百分比或部分。該距離臨限值可根據系統的靜態或動態參數，而由系統所決定。一旦該衍生被接受，候選元件就由重建程式及對一或多個主要資料元件的參照之組合所重制

定及置換。在提取資料中被建立用於該候選元件的條目係由該衍生所置換，亦即，它係由此係衍生物元件的指示，伴隨有重建程式加上對包含在該衍生中之一或多個主要資料元件的參照所置換(操作218)。另一方面，若用於最佳衍生之距離超過距離臨限值時(操作214的“否”分支)，則沒有可能的衍生物將被接受。在該情況中，該候選元件可被配置及進入到篩內，作為新的主要資料元件，且在提取資料中被建立用於該候選元件的條目將變成對該新建立之主要資料元件的參照，而伴隨有此係衍生物元件的指示(操作216)。

【0097】最後，該處理可檢查是否有任何額外的候選元件(操作222)，且若具有更多的候選元件時，則回到操作204(操作222的“是”分支)，或若沒有更多的候選元件時，則終止該處理(操作222的“否”分支)。

【0098】各種方法可被採用以執行第2圖中之操作202，亦即，解析輸入資料且將其拆散成候選元件。分解演算需決定位元組流中之何處應插入分隔符以切片該流成為候選元件。可能的技術包含(但不受限於)拆散該流成為固定大小區塊(諸如4096位元組的頁面)，或施加指紋圖譜之方法(諸如施加隨機的質多項式至輸入流的子字串)用以在資料流中設置合適的指紋圖譜而變成元件的邊界(此技術可導致可變大小元件)，或解析該輸入以偵測標頭或某些預宣告結構並根據此結構而描繪元件。該輸入可被解析以偵測透過輪廓所宣告的某一結構。該輸入可被解析以偵

測資料中之預宣告圖案、語法或正規表示的存在。一旦已識別出資料中之兩個接連的分隔符，候選元件就被產生(該候選元件係設置在兩個接連的分隔符之間的資料)，且被提交至主要資料篩，用於內容相關查找。若建立出可變大小元件時，則候選元件的長度需被指明及登載為伴隨有候選元件之元資料。

【0099】主要資料篩的一種重要功能在於根據所提交至它的候選元件而提供內容相關查找，且快速提供一或小的組之主要資料元件，其中主要資料元件自候選元件可以利用用以指明衍生所需的最小儲存之處取得。此係給定大的資料集之一大難題。給定甚至具有千個位元組大小之元件的萬億個位元組之資料，存在有數十億個元件要自該資料搜尋及選擇。在更大的資料集上，該問題甚至更為嚴重。為了要能快速提供小的組之合適的主要資料元件，使用合適的技術以組織及排列該等元件，且接著，在該等元件的該組織內偵測相似性及可取得性，就變得重要。

【0100】在篩中之條目可根據各元件(亦即，主要資料元件)的值而被排列，以致使所有的條目可藉由值而以上升或下降的順序配置。選擇性地，它們可沿著其係根據元件中之某些欄位的值之主軸，其次藉由使用該元件之內容的剩餘部分之從屬軸來排列。在此情況中，欄位係來自元件的內容之一組相連的位元組。欄位可藉由對元件的內容施加指紋圖譜之方法來設置，以致使指紋圖譜的位置識別欄位的位置。選擇性地，在元件的內容之內的某些固定

之偏移量可被選擇以設置欄位。其它的方法亦可被採用以設置欄位，包含但不限於解析元件以偵測某一宣告的結構並在該結構之內設置欄位。

【0101】在又另一形式的組織中，在元件內的某些欄位或欄位的組合可被視為維度，以致使該等維度隨後之各元件的內容之剩餘部分的串聯可被用來排列和組織資料元件。大體上，在欄位與維度之間的對應關係或映射可係任意地複雜。例如，在某些實施例中，正好一個欄位可映射至正好一個維度。在其它實施例中，例如，F1、F2及F3之多個欄位的組合可映射至維度。欄位的組合可藉由串聯兩個欄位，或藉由施加任何其它合適的功能至它們來達成。重要的需求在於，被用來組織元件之元件的欄位、維度，及其餘內容的配置必須使所有的主要資料元件變成能由它們的內容所唯一識別，且能在篩中被排列。

【0102】在又一實施例中，某些合適的函數(諸如代數或算術轉換)可被施加到元件，其中該函數具有函數的結果唯一地識別每個元件的屬性。在一個這種實施例中，每個元件除以質多項式或一些選定的數字或值，並且除法的結果(其包含商和餘數對)被用作用於組織和排序主要資料篩中的元件的函數。例如，包含餘數的位元可以形成函數的結果的前導位元組，其後是包含商的位元。或者，包含商的位元可以被用來形成函數的結果的前導位元組，其後是包含餘數的位元。對於用於分割輸入元件的給定除數，商和餘數對將唯一地識別該元件，因此此對可被用於

形成用於組織和排序主要資料篩中的元件的函數的結果。藉由將此函數施加於每個元件，可以根據函數的結果來在篩中組織主要資料元件。所述函數仍將唯一識別每個主要資料元件，並將提供一種替代方法來對主要資料篩中的主要資料元件進行分類和組織。

【0103】在又一個實施例中，可以對元件的每個欄位施加某些合適的函數(諸如代數或算術轉換)，其中函數具有函數的結果唯一地識別該欄位的屬性。例如，可以對於每個元件的內容的連續欄位或連續部分執行諸如除以合適的多項式或數字或值的函數，使得可以使用連續函數的結果的連接來排序和組織主要資料篩中的元件。請注意，對於每個欄位，可以使用不同的多項式來進行除法。每個函數都會為該部分或欄位提供來自由除法運算發出的商和餘數之位元的適當順序的串接。每個主要資料元件可以藉由使用施加於元件欄位的函數的此串接來在篩中排序和組織。函數的串接仍將唯一識別每個主要資料元件，並將提供一種替代方法來對主要資料篩中的主要資料元件進行分類和組織。

【0104】在一些實施例中，元件的內容可被表示成如下之表達式：元件=標頭.*sig1.*sig2.*...sigI.*...sigN.*尾部，其中“標頭(Head)”係包含元件之前導位元組的位元組序列，“尾部(Tail)”係包含元件之結束位元組的位元組序列，以及“sig1”、“sig2”、“sigI”及“sigN”係賦予元件特徵之元件內容本體內的某些長度之位元組的各種特徵碼或圖

案或正規表示或序列。在各種特徵碼之間的“*”表示係通配符表示，其係允許除了在“*”表示之後的特徵碼外之任何數目的任何值之插入位元組的正規表達符號。在一些實施例中，N元組(sig1, sig2, ... sigI, ...sigN)係稱作元件的骨幹資料結構或骨架，且可被視為元件之縮減的及根本的子集或本質。在其它的實施例中，(N+2)元組(Head, sig1, sig2, ... sigI,... sigN, Tail)被稱作元件的骨幹資料結構或骨架。選擇性地，可採用N+1元組，其包含Head或Tail，而伴隨有該等特徵碼的剩餘部分。

【0105】 可將指紋圖譜之方法施加至元件的內容，以決定該元件內容之內的骨幹資料結構之各種組件(或特徵碼)的位置。選擇性地，在元件的內容之內的某些固定之偏移量可被選擇以設置組件。其它的方法亦可被採用以設置骨幹資料結構之組件，包含但不限於解析元件以偵測某一宣告的結構並在該結構之內設置組件。主要資料元件可根據骨幹資料結構而在篩中被排列。換言之，元件的骨幹資料結構之各種組件可被視為維度，以致使這種維度隨後之各元件的內容之剩餘部分的串聯可被用來在篩中排列和組織主要資料元件。

【0106】 一些實施例將輸入資料分解成候選元件，其中各候選元件的大小係實質大於用以存取全域資料集中之所有該等元件所需之參照的大小。關於被拆散成該等資料塊之資料(且其將以內容相關方式存取)的一觀察在於，相對於資料塊可指明之全部可能的值，實際資料係非常稀疏

的。例如，考慮1千萬億兆個位元組的資料集。用以定址該資料集中之每個位元組大約需要70個位元。就一個大小為128個位元組(1024個位元)的資料塊而言，在1千萬億兆個位元組的資料集中，約具有 2^{63} 個資料塊，以致需要63個位元(少於8位元組)以定址所有該等資料塊。應注意的是，1024個位元的元件或資料塊可具有 2^{1024} 個可能的值之一者，而在資料集中之給定資料塊的實際值之數目係最大 2^{63} (若所有該等資料塊係不同時)。此指示的是，相對於可由元件之內容所達到或命名的值之數目，實際資料係極稀疏。此致能樹狀結構的使用，該樹狀結構係非常適合於以可致能有效率之內容為主查找的方式組織非常稀疏之資料，允許新的元件被有效率地添加至該樹狀結構，以及就用於該樹狀結構本身所需之增量的儲存而言，係經濟有效的。雖然在1千萬億兆個位元組的資料集中僅具有 2^{63} 個資料塊，因而僅需要63個不同位元的資訊以告訴它們分開，但相關聯之不同的位元可能散佈在元件之整個的1024個位元之範圍，且發生在各元件之不同位置處。因此，僅從內容檢驗固定的63個位元係不夠的，反而是，元件的整個內容需參與該等元件的排序，具體地，需參與其中對資料集中之任何及每個元件提供真正內容相關存取的解決方案。在資料提取(Data Distillation™)架構中，希望能夠在被用來排列及組織資料的架構內偵測出可取得性。應謹記所有上述的，根據內容的樹狀結構(當檢驗更多的內容時，其逐步區分資料)係用以排列及區分所分解之資料集中的所

有元件之合適組織。該結構提供許多中間層次的子樹狀物，其可被視為衍生物元件之群組或具有相似衍生性質的元件之群組。該結構可以用顯示各子樹狀物之特徵的元資料，或以顯示資料的各元件之特徵的元資料階層地增加。該結構可有效地連繫其所包含之整個資料的組成，包含資料中之實際值的密度、鄰近度及分布。

【0107】 一些實施例以樹狀形式來組織篩中的主要資料元件。各主要資料元件具有不同的“名稱”，其係由該主要資料元件的整個內容所建構。此名稱係設計成為足以唯一地識別該主要資料元件，且相對於樹狀物中的所有其它元件而區分它。有其中該名稱可由主要資料元件之內容所建構的若干方式。該名稱可單純地包含主要資料元件的所有位元組，而該等位元組以與它們在主要資料元件中所存在的相同順序呈現在該名稱中。在另一實施例中，被稱作維度的某些欄位或欄位的組合(其中欄位和維度係如稍早所敘述的)係使用來形成名稱的前導位元組，而主要資料元件之內容的剩餘部分則形成名稱的剩餘部分，以致使主要資料元件的整個內容參與而建立出元件之完整且唯一的名稱。在又另一實施例中，元件之骨幹資料結構的欄位被選擇作為維度(其中欄位和維度係如稍早所敘述的)，且被用來形成名稱的前導位元組，而主要資料元件之內容的剩餘部分則形成名稱的剩餘部分，以致使主要資料元件的整個內容參與而建立出完整且唯一的元件名稱。

【0108】 在一些實施例中，元件的名稱可以藉由對元

件執行代數或算術轉換來計算，同時保留每個名稱唯一識別每個元件的屬性。在一個這種實施例中，每個元件除以質多項式或一些選定的數字或值，並且除法的結果(其為商和餘數對)形成元件的名稱。例如，包含餘數的位元可以形成名稱的前導位元組，其後是包含商的位元。或者，包含商的位元可以被用來形成名稱的前導位元組，其後是包含餘數的位元。對於用於分割輸入元件的給定除數，商和餘數對將唯一地識別該元件，因此此對可被用於形成每個元件的名稱。使用名稱的此公式，可以根據其名稱來在篩中組織主要資料元件。所述名稱仍將唯一識別每個主要資料元件，並將提供一種替代方法來對主要資料篩中的主要資料元件進行分類和組織。

【0109】 在另一實施例中，可以採用產生名稱的這種方法(其關於除法和解壓縮商數/餘數對)的變體，其中可以對於每個元件的內容的連續欄位或連續部分執行除以合適的多項式或數字或值的除法，針對每個元件產生名稱的連續部分(每個部分為針對該部分或欄位，來自由除法運算發出的商和餘數之位元的適當順序的串接)。請注意，對於每個欄位，可以使用不同的多項式來進行除法。使用名稱的此公式，可以根據其名稱來在篩中組織主要資料元件。所述名稱仍將唯一識別每個主要資料元件，並將提供一種替代方法來對主要資料篩中的主要資料元件進行分類和組織。

【0110】 各主要資料元件的名稱係使用來在樹狀物中

排列及組織主要資料元件。對於最實際的資料組，即使在大小上係非常大之該等者(諸如1千萬億兆個位元組，包含比如說，4KB大小的 2^{58} 個元件)，所期望的是，名稱的該等位元組之小的子集將時常用以排序及排列樹狀物中之大部分的主要資料元件。

【0111】第3A、3B、3C、3D及3E圖顯示根據在此所敘述的一些實施例之不同的資料組織系統，其可使用以根據主要資料元件名稱來組織它們。

【0112】第3A圖顯示前綴樹(trie)資料結構，其中主要資料元件係根據來自各主要資料元件的名稱之連續位元組的值來組織成為逐步更小的群組。在第3A圖中所示的範例中，各主要資料元件具有不同的名稱，其係由主要資料元件之整個內容所建構，且此名稱單純地包含主要資料元件的所有位元組，而該等位元組以與它們在主要資料元件中所存在的相同順序呈現在該名稱中。該前綴樹之根節點表示所有的主要資料元件。該前綴樹之其它節點表示主要資料元件的子集或群組。從前綴樹的根節點或第一層次開始(在第3A圖中被標記為根302)，主要資料元件係根據它們名稱之最高有效位元組的值而被群組化成子樹狀物(在第3A圖中被標記為N1)。在它們名稱之最高有效位元組中具有相同值的所有主要資料元件將被一起群組化成共同子樹狀物，且由該值所表示之鏈路將存在於自根節點至表示該子樹狀物的節點。例如，在第3A圖中，節點303表示主要資料元件的子樹狀物或群組，其各在它們個別名稱之它

們的最高有效位元組 N1 中具有相同的值 2。在第 3A 圖中，此群組包含主要資料元件 305、306 及 307。

【0113】在前綴樹的第二層次，各主要資料元件之名稱的第二高有效位元組係使用來進一步劃分該主要資料元件之各群組成為更小的子群組。例如，在第 3A 圖中，由節點 303 所表示之主要資料元件的群組係使用第二高有效位元組 N2 來進一步劃分成子群組。節點 304 表示主要資料元件的子群組，其在它們的最高有效位元組 N1 中具有值 2，且亦在它們個別名稱之它們的第二高有效位元組 N2 中具有值 1。此群組包含主要資料元件 305 及 306。

【0114】分部在該前綴樹的各層次處繼續處理，而建立出從父節點到各子節點的鏈路，其中子節點代表由父節點所表示之主要資料元件的子集。此處理持續著，直到該前綴樹的葉狀物處僅具有個別的主要資料元件為止。葉狀節點表示葉的群組。在第 3A 圖中，節點 304 係葉狀節點。由節點 304 所表示之主要資料元件的群組包含主要資料元件 305 及 306。在第 3A 圖中，此群組係使用它們名稱之第三高有效位元組來進一步劃分成個別的主要資料元件 305 及 306。N3=3 的值導致主要資料元件 305，而 N3=5 的值導致主要資料元件 306。在此範例中，在它們整個名稱中，僅 3 個有效位元組就足以完全地識別主要資料元件 305 及 306。同樣地，僅從名稱的兩個有效位元組就足以識別主要資料元件 307。

【0115】此範例描繪僅只名稱之位元組的子集如何在

給定混合的主要資料元件中用以識別樹狀物中之主要資料元件，且整個名稱並不需要到達單獨的主要資料元件。再者，主要資料元件或主要資料元件的群組可各自要求不同數目的位元組，以便能獨特地識別它們。因此，自根節點至主要資料元件之前綴樹的深度可自一主要資料元件變化至另一者。此外，各節點可具有下降至底下之子樹狀物的不同數目之鏈路。

【0116】在這種前綴樹中，各節點具有名稱，其包含指明如何到達此節點之位元組的序列。例如，用於節點304的名稱係“21”。此外，來自元件的名稱之其唯一識別樹狀物中的元件之目前分布中的元件之位元組的子集是從根節點到此主要資料元件的“路徑”。例如，在第3A圖中，具有213之值的路徑301識別主要資料元件305。

【0117】在此所敘述之前綴樹結構可建立出深厚的樹狀物(亦即，具有許多層次的樹狀物)，因為在樹狀物中的元件名稱之每個不同的位元組將添加一層次之深度至該前綴樹。

【0118】應注意的是，在第3A至3E圖中之樹狀資料結構已從左邊予以繪製到右邊。因此，當我們從圖式的左側移動到圖式的右側時，我們將從樹狀物的高層次移動到樹狀物的低層次。在給定的節點底下(亦即，朝向第3A至3E圖中之給定節點的右邊)，對於藉由來自名稱之不同位元組的確定值而被選擇的任何子節點，駐存在該子節點底下之子樹狀物中的所有元件將在元件名稱中之該對應位元

組中具有相同的值。

【0119】我們現在敘述在給定輸入候選元件時之用於前綴樹結構的內容相關查找之方法。此方法包含使用候選元件的名稱之前綴樹結構的導航，其次藉由隨後的分析和篩查以決定應回報何者作為整體之內容相關查找的結果。換言之，前綴樹導航處理回報第一輸出，且然後，分析和篩查被執行於該輸出上，而決定整體之內容相關查找的結果。

【0120】在開始該前綴樹導航處理中，來自候選元件的名稱之最高有效位元組的值將被用來選擇從根節點到隨後之節點的鏈路(由該值所表示)，該隨後之節點表示在該等主要資料元件之名稱的最高有效位元組中具有該相同值之主要資料元件的子樹狀物。自此節點出發，來自候選元件的名稱之第二位元組被檢驗，且由該值所表示的鏈路被選擇，而藉以推動更深(或更下方)一層次至該前綴樹之內，且選擇更小之子群組的主要資料元件，該更小之子群組的主要資料元件目前與該候選元件共有著來自名稱之至少兩個有效位元組。此處理持續著，直至到達單一的主要資料元件，或直到並無鏈路匹配於來自候選元件的名稱之對應位元組的值為止。在該等情形的任一者之下，若到達單一的主要資料元件時，則可將其回報當做前綴樹導航處理的輸出。若並未到達時，則一選擇例係報告“失敗”。另一選擇例係回報多個主要資料元件，而該多個主要資料元件係在被植根於其中該導航終止之節點處的子樹狀物中。

【0121】一旦該前綴樹導航處理已終止，其它的準則及要求就可被用來分析和篩查該前綴樹導航處理的輸出，而決定應回報何者作為該內容相關查找的結果。例如，當單一主要資料元件或多個主要資料元件係由前綴樹導航處理所回報時，則可具有額外的要求，其中在給予將被回報作為該內容相關查找的結果之資格前，它們與該候選元件的名稱共享某一最小數量的位元組(否則，該內容相關查找回報失敗)。篩查要求的另一範例可以是：若該前綴樹導航處理終止而未到達單一的主要資料元件，以致使多個主要資料元件(被植根於其中該前綴樹導航終止之節點處)被回報作為前綴樹導航處理的輸出時，則除非該等元件的數量小於某一特定的限制值，該等多個主要資料元件才將被給予回報作為該內容相關查找的結果之資格(否則，該內容相關查找回報失敗)。多個要求的組合可被採用以決定內容相關查找的結果。以這種方式，查找處理將報告“失敗”或回報單一主要資料元件，或者若非單一的主要資料元件時，則一組主要資料元件可能成為用以取得該候選元件之好的起點。

【0122】下文所描述的第3B至3E圖有關第3A圖中所描繪之樹狀資料結構的變化例和修正例。雖然該等變化例提供凌駕於第3A圖中所描繪之前綴樹資料結構的增進及優點，但用以導航該資料結構的處理係與上文參照第3A圖所敘述的處理相似。也就是說，在用於第3B至3E圖中所示之樹狀資料結構的導航終止，且隨後的分析和篩查被執行

以決定整體之內容相關查找的結果之後，整個處理回報失敗、單一的主要資料元件或可能成為用以取得該候選元件之好的起點之一組主要資料元件。

【0123】第3B圖顯示另一資料組織系統，其可被用來根據主要資料元件的名稱來組織它們。在第3B圖中所示的範例中，各主要資料元件具有不同的名稱，其係由主要資料元件之整個內容所建構，且此名稱單純地包含主要資料元件的所有位元組，而該等位元組以與它們在主要資料元件中所存在的相同順序呈現在該名稱中。第3B圖顯示更緊湊的結構，其中單一鏈路採用來自主要資料元件之名稱的多個位元組(而非第3A圖中之前綴樹中所使用的單一位元組)於下面的子樹狀物中，而建立出分部或下一層次的群組。從父節點到子節點的鏈路現在係由多個位元組所表示。進一步地，從任一給定的父節點起，各鏈路可採用不同數目的位元組，用以區分及識別與該鏈路相關聯的子樹狀物。例如，在第3B圖中，從根節點到節點308的鏈路係藉由來自名稱的4個位元組($N_1N_2N_3N_4=9845$)來區分，且同時，從根節點到節點309的鏈路係由來自名稱的3個位元組($N_1N_2N_3=347$)所區分。

【0124】應注意的是，在樹狀導航之期間(使用來自給定之候選元件的內容)，一旦到達樹狀物中之任一父節點，該樹狀導航處理就需確保的是，來自候選元件之名稱的足夠位元組被檢驗，以便明白地決定要選擇那一個鏈路。為了要選擇給定的鏈路，來自候選元件之名稱的該等

位元組必須與對該特殊鏈路之轉換所指示的所有位元組匹配。再次地，在該樹狀物中，樹狀物的各節點具有名稱，其包含位元組之序列，用以指明如何到達此節點。例如，節點309的名稱可係“347”，因為它代表具有其名稱之3個前導位元組為347的主要資料元件之群組(例如，元件311及312)。在使用具有名稱之前導3個位元組為347的候選元件之樹狀查找時，此資料圖案致使樹狀導航處理到達節點309，如第3B圖中所示。再次地，來自元件的名稱之其唯一識別樹狀物中的元件之目前分布中的元件之位元組的子集是從根節點到此主要資料元件的“路徑”。例如，在第3B圖中，位元組的序列3475導致主要資料元件312，且在該範例中所示之混合的主要資料元件中唯一地識別主要資料元件312。

【0125】對於多樣性和稀疏的資料，在第3B圖中的樹狀結構可提供比第3A圖的前綴樹結構更多的彈性和緊湊。

【0126】第3C圖顯示另一資料組織系統，其可被用來根據主要資料元件的名稱來組織它們。在第3C圖中所示的範例中，各主要資料元件具有不同的名稱，其係由主要資料元件之整個內容所建構，且此名稱單純地包含主要資料元件的所有位元組，而該等位元組以與它們在主要資料元件中所存在的相同順序呈現在該名稱中。第3C圖顯示另一變化例(對第3B圖中所描述的組織)，其藉由使用正規表示(其中必要且/或有用)，以指明來自主要資料元件的名稱之導致各種鏈路的值，而使樹狀物或子樹狀物中的群組元件

進一步緊湊化。該正規表示的使用允許在相同的子樹狀物下之對應位元組上，所共享相同表示之元件的有效率群組；此可接著由子樹狀物內之不同主要資料元件的更多局部之澄清所跟隨。此外，該正規表示的使用允許更緊湊之方式以描述用以映射元件至底下任一子樹狀物所需之位元組的值。此進一步降低用以指明樹狀物所需之位元組的數目。例如，正規表示318指明28個連續“F”的圖案；若在樹狀導航的期間跟隨著此鏈路，則可到達元件314，其包含具有按正規表示318之28個連續“F”的圖案320。同樣地，到達元件316的路徑具有鏈路或分支，其使用指明具有16個連續“0”之圖案的正規表示。對於該樹狀物，樹狀導航處理需偵測和執行該等正規表示，以便決定要選擇那一個鏈路。

【0127】第3D圖顯示另一資料組織系統，其可被用來根據主要資料元件的名稱來組織它們。在第3D圖中所示的範例中，各主要資料元件具有不同的名稱，其係由主要資料元件之整個內容所建構。指紋圖譜之方法係施加至各元件，用以識別其中包含評估為選定指紋圖譜之欄位的位置。在元件中所發現的第一指紋圖譜之位置處的欄位係視為維度，且來自此欄位之某一數目的位元組(比方說， x 個位元組，其中 x 係顯著地小於該元件中之位元組的數目)被提取及使用作為元件之名稱的前導位元組，而名稱之位元組的剩餘部分包含主要資料元件之位元組的剩餘部分，並以它們在主要資料元件中所存在之相同的循環順序呈現。

此名稱係使用來在樹狀物中組織主要資料元件。在此範例中，當在元件中並未偵測出指紋圖譜時，則名稱係藉由單純地使用元件的所有位元組而以其中它們在元件中所存在的順序制定。個別的子樹狀物(由並未發現指紋圖譜的指示所表示)根據元件之它們的名稱而保持和組織所有該等元件。

【0128】例如，如第3D圖中所示，指紋圖譜技術可被施加至元件338(其包含 t 個位元組的資料，亦即， $B_1B_2B_3...B_t$)，用以在位元組 B_{i+1} 獲得指紋圖譜位置“指紋圖譜1”，而識別將被選定為“維度1”的欄位。接著，從“指紋圖譜1”所識別之位置起的 x 個位元組可被提取以形成“維度1”，且該等 x 個位元組可被使用作為第3D圖中之各元件的名稱之前導位元組 $N_1N_2...N_x$ 。之後，來自元件338之剩餘部分的 $t-x$ 個位元組(自 B_{i+x+1} 開始，且稍後繞回至 $B_1B_2B_3...B_i$)被串聯及使用作為名稱之剩餘部分的位元組 $N_{x+1}N_{x+2}...N_t$ 。當在元件中並未發現指紋圖譜時，則名稱 $N_1N_2...N_t$ 係單純地來自元件338的 $B_1B_2B_3...B_t$ 。主要資料元件係使用它們的名稱而在樹狀物中被排序和組織。例如，主要資料元件(PDE)330係在使用路徑13654...06而遍歷樹狀物的兩層次之後識別及到達，其中該等位元組13654...0係 $N_1N_2...N_x$ ，其係來自維度1的位元組。沿著鏈路334而由根所到達之節點335處的個別子樹狀物(由並未發現指紋圖譜的指示所表示)，保持及組織其內容並未評估為選定指紋圖譜的所有主要資料元件。因此，在此組織中，例如，

鏈路 336 之某些鏈路可使用其包含以與元件中相同之順序所呈現的元件之位元組的名稱，以組織元件，而例如，鏈路 340 之其它鏈路可使用其係使用指紋圖譜而被制定的名稱來組織元件。

【0129】在接收到候選元件時，處理施加上述相同之技術以決定該候選元件的名稱，且使用此名稱而導航樹狀物，以供內容相關查找之用。因此，相同且一致的處理被施加至主要資料元件(在將它們安裝至樹狀物之內時)，及至候選元件(在從解析及分解器接收到它們時)，以便建立出它們的名稱。樹狀導航處理使用候選元件的名稱以導航樹狀物。在此實施例中，若在候選元件中並未發現指紋圖譜時，該樹狀導航處理向下導航至子樹狀物，其中該子樹狀物組織及包含其內容並未評估為指紋圖譜的主要資料元件。

【0130】第 3E 圖顯示另一資料組織系統，其可被用來根據主要資料元件的名稱來組織它們。在第 3E 圖中所示的範例中，各主要資料元件具有不同的名稱，其係由主要資料元件之整個內容所建構。指紋圖譜之方法係施加至各元件，用以識別其中包含評估為兩個指紋圖譜之任一者的內容之欄位的位置。在元件中之第一指紋圖譜(在第 3E 圖中之指紋圖譜 1)的第一出現之位置處的欄位係視為第一維度(維度 1)，以及設置在第二指紋圖譜(在第 3E 圖中之指紋圖譜 2)的第一出現處的欄位係視為第二維度(維度 2)。用以搜索兩個不同指紋圖譜之指紋圖譜的使用導致四種可能的情

況：(1)該兩個指紋圖譜均在元件中被發現；(2)指紋圖譜1被發現，但指紋圖譜2並未被發現；(3)指紋圖譜2被發現，但指紋圖譜1並未被發現；以及(4)沒有指紋圖譜被發現。主要資料元件可被群組成為對應該等情況之各者的四個子樹狀物。在第3E圖中，“FP1”表示存在指紋圖譜1，“FP2”表示存在指紋圖譜2，“ $\sim$ FP1”表示不存在指紋圖譜1，以及“ $\sim$ FP2”表示不存在指紋圖譜2。

【0131】對於該四種情況之各者，元件的名稱係如下地建立出：(1)當該兩個指紋圖譜均被發現時，從“指紋圖譜1”所識別之位置起的 x 個位元組可被提取以形成“維度1”，且從“指紋圖譜2”所識別之位置起的 y 個位元組可被提取以形成“維度2”，以及該等 $x+y$ 個位元組可被使用作為第3E圖中之各該元件的名稱之前導位元組 $N_1N_2\ldots N_{x+y}$ 。之後，來自元件348之剩餘部分的 $t-(x+y)$ 個位元組係以循環方式提取(在從第一維度的該等位元組之後開始)，且被串聯及使用作為名稱之剩餘部分的位元組 $N_{x+y+1}N_{x+y+2}\ldots N_t$ 。(2)當非指紋圖譜2，而是指紋圖譜1被發現時，從“指紋圖譜1”所識別之位置起的 x 個位元組可被提取以形成前導大小，且該等 x 個位元組可被使用作為各該元件的名稱之前導位元組 $N_1N_2\ldots N_x$ 。之後，來自元件348之剩餘部分的 $t-x$ 個位元組(自 B_{i+x+1} 開始，且稍後繞回至 $B_1B_2B_3\ldots B_i$)被串聯及使用作為名稱之剩餘部分的位元組 $N_{x+1}N_{x+2}\ldots N_t$ 。(3)當非指紋圖譜1，而是指紋圖譜2被發現時，從“指紋圖譜2”所識別之位置起的 y 個位元組可被提取以形成前導維度，

且該等 y 個位元組可被使用作為各該元件的名稱之前導位元組 $N_1 N_2 \dots N_y$ 。之後，來自元件 348 之剩餘部分的 $t-y$ 個位元組 (自 B_{j+x+1} 開始，且稍後繞回至 $B_1 B_2 B_3 \dots B_j$) 被串聯及使用作為名稱之剩餘部分的位元組 $N_{y+1} N_{y+2} \dots N_t$ 。(4) 當在元件中並未發現指紋圖譜時，則名稱 $N_1 N_2 \dots N_t$ 係單純地來自元件 348 的 $B_1 B_2 B_3 \dots B_t$ 。因此，個別的子樹狀物存在以供該等四種情況之用。用以提取用於元件 348 之名稱 ($N_1 N_2 N_3 \dots N_t$) 的處理可針對該四種情況而被概述如下：

(1) 指紋圖譜 1 及指紋圖譜 2 二者均被發現：

$N_1 - N_x \leftarrow B_{i+1} - B_{i+x}$ = 來自維度 1 之 x 個位元組

$N_{x+1} - N_{x+y} \leftarrow B_{j+1} - B_{j+y}$ = 來自維度 2 之 y 個位元組

$N_{x+y+1} \dots N_t$ = 剩餘部分的位元組 (來自大小 t 個位元組之候選元件) = $B_{i+x+1} B_{i+x+2} B_{i+x+3} \dots B_j B_{j+y+1} B_{j+y+2} B_{j+y+3} \dots B_t B_1 B_2 B_3 \dots B_i$

(2) 指紋圖譜 1 被發現，但指紋圖譜 2 並未被發現：

$N_1 - N_x \leftarrow B_{i+1} - B_{i+x}$ = 來自維度 1 之 x 個位元組

$N_{x+1} \dots N_t$ = 剩餘部分的位元組 (來自維度 t 個位元組之候選元件) = $B_{i+x+1} B_{i+x+2} B_{i+x+3} \dots B_t B_1 B_2 B_3 \dots B_i$

(3) 指紋圖譜 2 被發現，但指紋圖譜 1 並未被發現：

$N_1 - N_y \leftarrow B_{j+1} - B_{j+y}$ = 來自維度 2 之 y 個位元組

$N_{y+1} \dots N_t$ = 剩餘部分的位元組 (來自維度 t 個位元組之候選元件) = $B_{j+y+1} B_{j+y+2} B_{j+y+3} \dots B_t B_1 B_2 B_3 \dots B_j$

(4) 沒有指紋圖譜被發現

$N_1 - N_x \leftarrow B_1 - B_t$

【0132】在接收到候選元件時，處理施加上述相同之技術以決定該候選元件的名稱。在此實施例中，上述名稱結構的4個方法(根據指紋圖譜1及指紋圖譜2是否被發現)係施加至候選元件，正如當它們進入至篩之中時，它們係用於主要資料元件。因此，相同且一致的處理被施加至主要資料元件(在將它們安裝至樹狀物之內時)，及至候選元件(在從解析及分解器接收到它們時)，以便建立出它們的名稱。樹狀導航處理使用候選元件的名稱而導航樹狀物，以供內容相關查找之用。

【0133】若內容相關查找成功時，則將在特定維度之位置處產生具有與候選元件相同圖案的主要資料元件。例如，若在候選元件中發現該等指紋圖譜二者時，則該樹狀導航處理將從根節點開始而向下到樹狀物之鏈路354。若候選元件具有圖案“99...3”作為‘維度1’以及圖案“7...5”作為‘維度2’時，該樹狀導航處理將到達節點334。這到達了包含兩個主要資料元件(PDE 352及PDE 353)的子樹狀物，該兩個主要資料元件可能係用於衍生之目標。額外的分析及篩查係執行(藉由首先檢驗元資料，且視需要地，藉由隨後提取及檢驗實際的主要資料元件)，用以決定那一個主要資料元件係最佳地適用於該衍生。因而，在此所敘述之實施例將識別可被使用於篩中的各種樹狀結構。該等結構或其變化例的組合可被採用以組織該等主要資料元件。某些實施例以樹狀形式組織該等主要資料元件，其中元件的整個內容係使用作為該元件的名稱。然而，其中位元組

呈現在元件的名稱中之序列無需一定要是其中該等位元組出現在該元件之中的序列。元件的某些欄位係提取作為維度且被用來形成名稱的前導位元組，以及元件之剩餘部分的位元組則補足名稱的剩餘部分。使用該等名稱，該等元件可以用樹狀形式排列於篩中。名稱的前導數字係使用來區分樹狀物之較高的分支(或鏈路)，以及剩餘部分的數字則被用來逐步地區分該樹狀物的所有分支(或鏈路)。樹狀物的各節點可具有從該節點所發出之不同數目的鏈路。此外，來自節點之各鏈路可藉由不同數目的位元組來區分和表示，以及該等位元組之說明可透過用以表達它們的規格之正規表示及其它強大方式的使用而被完成。所有該等特徵導致緊湊的樹狀結構。在樹狀物的葉狀節點處駐存著對個別之主要資料元件的參照。

【0134】在一個實施例中，指紋圖譜的方法可被施加至包含主要資料元件的位元組。駐存在由指紋圖譜所識別之位置處的若干位元組可被用來組成元件名稱的組件。一或多個組件可被結合以提供維度。多個指紋圖譜可被用來識別多個維度。這種維度係串聯且使用作為元件之名稱的前導位元組，而元件之剩餘部分的位元組構成元件之剩餘部分的名稱。因為該等維度係設置在由指紋圖譜所識別之位置處，所以可增加名稱將由來自各元件之一致內容所形成的可能性。在由指紋圖譜所設置的欄位處具有相同值之內容的元件，將沿著樹狀物之相同的分支而被群組在一起。以這種方式，相似的元件將在樹狀資料結構中被群組

在一起。在它們之中所發現之不具有指紋圖譜的元件可使用它們名稱之替代的制定來一起群組在個別的子樹狀物中。

【0135】 在一個實施例中，指紋圖譜的方法可被施加至元件之內容，用以決定元件之內容內的骨幹資料結構(如稍早所敘述)之各種組件(或特徵碼)的位置。選擇性地，在元件之內容內之某些固定的偏移可被選擇以設置組件。其它的方法亦可被採用以設置元件之骨幹資料結構的組件包含(但不受限於)解析元件以偵測出某一宣告的結構，且設置組件於該結構之內。元件之骨幹資料結構的各種組件可被視為維度，以致使各元件之剩餘部分的內容所跟隨之該等維度的串聯被用來建立出各元件的名稱。該名稱係使用來在樹狀物中排列及組織該等主要資料元件。

【0136】 在另一實施例中，元件被解析以便偵測出元件中的某一結構。在此結構中之某些欄位係識別作為維度。多個這種維度係串聯且使用作為名稱的前導位元組，而元件之剩餘部分的位元組構成元件之剩餘部分的名稱。因為該等維度係設置在藉由解析元件及偵測其結構來識別的位置處，所以可增加名稱將由來自各元件之一致內容所形成的可能性。在藉由該解析來設置的欄位處具有相同值之內容的元件，將沿著樹狀物之相同的分支而被群組在一起。以這種方式，相似的元件將在樹狀資料結構中被群組在一起。

【0137】 在某些實施例中，在樹狀資料結構中的各節

點包含自描述規格。樹狀節點具有一或多個子。各子條目包含在對該子之鏈路上的不同位元組上的資訊，及對該子節點的參照。子節點可係樹狀節點或葉狀節點。第3F圖呈現根據在此所敘述的一些實施例之自描述樹狀節點資料結構。在第3F圖中所示的樹狀節點資料結構指明(A)有關從根節點到此樹狀節點之路徑的資訊，包含所有以下之組件，或以下之組件的子集：用以到達此樹狀節點之來自名稱的實際位元組序列，用以從根節點到達此節點所消耗之名稱的位元組數目，所消耗之此位元組數目是否大於某一預指明臨限值的指示，以及描述對此節點的路徑，且有用於樹狀物之內容相關搜尋及有用於關於樹狀物之結構的決定之其它的元資料；(B)該節點所具有的子數目；(C)對於各子(其中各子對應樹狀物的分支)，將指明(1)子ID，(2)來自名稱以後的位元組所需之不同位元組的數目，以便向下轉換至樹狀物之此鏈路，(3)用以記下此鏈路之來自名稱的位元組之實際值的規格，以及(4)對該子節點的參照。

【0138】第3G圖呈現根據在此所敘述的一些實施例之自描述葉狀節點資料結構。葉狀節點具有一或多個子。各子係對主要資料元件之鏈路。各子條目包含在對該主要資料元件之鏈路上的不同位元組上的資訊、對該主要資料元件的參照，以及有關該主要資料元件之複製品和衍生物及其它的元資料的計數。在第3G圖中所示的葉狀節點資料結構指明(A)有關從根節點到此葉狀節點之路徑的資訊，包

含所有以下之組件，或以下之組件的子集：用以到達此葉狀節點之來自名稱的實際位元組序列，用以從根節點到達此節點所消耗之名稱的位元組數目，所消耗之此位元組數目是否大於某一預指明臨限值的指示，以及描述對此節點的路徑，且有用於樹狀物之內容相關搜尋及有用於關於樹狀物之結構的決定之其它的元資料；(B)該節點所具有的子數目；(C)對於各子(其中各子對應該葉狀節點之下的主要資料元件)，將指明(1)子ID，(2)來自名稱以後的位元組所需之不同位元組的數目，以便向下轉換樹狀物之此鏈路至主要資料元件，(3)用以記下此分支之來自名稱的位元組之實際值的規格，(4)對在樹狀物之此路徑上終止該樹狀物的主要資料元件的參照，(5)有多少複製品和衍生物正指向此主要資料元件的計數(此係使用來確定當刪除儲存系統中之資料時，條目是否可自篩刪除)，以及(6)用於該主要資料元件的其它元資料，包含主要資料元件之大小，等等。

【0139】為了要增加其中新的主要資料元件被安裝至樹狀物之內的效率，某些實施例合併額外的欄位至葉狀節點資料結構之內，以供被保存在該樹狀物之該葉狀節點處的各主要資料元件之用。應注意的是，當新的元件必須被插入至樹狀物之內時，在該樹狀物中可能需要該等主要資料元件之各者的名稱或內容之額外的位元組，以便決定要在子樹狀物中的何處插入該新的元件，或者是否要觸發該子樹狀物的進一步分區。對於該等額外位元組的需求，可

要求提取該等主要資料元件的若干者，以便相對於該新的元件而提取該等元件的各者之相關聯的不同位元組。為了要降低及最佳化(且在大多數的情況中，完全地消除)此任務所需之IO的數目，在葉狀節點中的資料結構包含來自各主要資料元件的名稱之一定數目的額外位元組於該葉狀節點之下。該等額外的位元組係稱作導航預看位元組，且相對於新的輸入元件而協助排序該等主要資料元件。用於給定之主要資料元件的導航預看位元組係在安裝該主要資料元件至節內時，被安裝至葉狀節點結構內。將被保留用於此目的之位元組的數目可使用各種準則來靜態或動態地選擇，包含，所包含之子樹狀物的深度及在該子樹狀物中之主要資料元件的密度。例如，對於將被安裝在樹狀物中之淺的層次處之主要資料元件，解決方法可添加比用於駐存在很深的樹狀物中之主要資料元件更長的導航預看欄位。此外，當新的元件將被安裝至節之內，且若已有許多主要資料元件在現有的目標子樹狀物之中時(增加即將來臨之重新分區的可能性)，則當新的主要資料元件將被安裝至該子樹狀物之內時，額外的導航預看位元組可被保留用於該新的主要資料元件。

【0140】第3H圖呈現用於葉狀節點的葉狀節點資料結構，其包含導航預看欄位。此資料結構指明(A)有關從根節點到此葉狀節點之路徑的資訊，包含所有以下之組件，或以下之組件的子集：用以到達此葉狀節點之來自名稱的實際位元組序列，用以從根節點到達此節點所消耗之名稱

的位元組數目，所消耗之此位元組數目是否大於某一預指明臨限值的指示，以及描述對此節點的路徑，且有用於樹狀物之內容相關搜尋及有用於關於樹狀物之結構的決定之其它的元資料；(B)該節點所具有的子數目；(C)對於各子(其中各子對應該葉狀節點之下的主要資料元件)，將指明(1)子ID，(2)來自名稱以後的位元組所需之不同位元組的數目，以便向下轉換樹狀物之此鏈路至主要資料元件，(3)用以記下此分支之來自名稱的位元組之實際值的規格，(4)對在樹狀物之此路徑上終止該樹狀物的主要資料元件的參照，(5)指明有多少導航預看的位元組被保留用於主要資料元件的導航預看欄位，及該等位元組的實際值，(6)有多少複製品和衍生物正指向此主要資料元件的計數(此係使用來確定當刪除儲存系統中之資料時，條目是否可自篩刪除)，以及(7)用於該主要資料元件的其它元資料，包含主要資料元件之大小，等等。

【0141】在某些實施例中，樹狀物的各種分支係使用來將各種資料元件映射成群組或範圍，該等群組或範圍係藉由解讀沿著鏈路通往子子樹狀物之不同的位元組作為分界符而被形成。在該子子樹狀物中的所有元件將變成，使得元件中之對應位元組的值小於或等於被指明用於對特殊之子子樹狀物的鏈路之不同位元組的值。因此，各子樹狀物現將表示其值落在特定範圍內之元件的群組。在給定的子樹狀物內，樹狀物的每個隨後層次將逐步地將該組元件劃分成較小的範圍。此實施例提供不同的解讀至第3F圖中

所示之自描述樹狀節點結構。在第3F圖中之N個子係由樹狀節點資料結構中之它們的不同位元組的值所排列，且表示非重疊範圍的排列順序。對於N個節點，具有N+1個範圍---最低的或第一個範圍包含小於或等於最小的條目值，以及第N+1個範圍包含大於第N個條目之值。該第N+1個範圍將被視為超出範圍，以致使N個鏈路通往在下面的N個子樹狀物或範圍。

【0142】例如，在第3F圖中，子1界定最低的範圍且使用(值abef12d6743a的)6個位元組以區分其範圍---用於子1的範圍係自00000000至abef12d6743a。若候選元件之對應的6個位元組係落在此範圍之內，包含最終值時，則此子之鏈路將被選定。若候選元件之對應的6個前導位元組係大於範圍分界符abef12d6743a時，則子1將不被選擇。為了要檢驗該候選者是否落在用於子2的範圍之內，兩種情形必須被滿足---首先，該候選者必須在用於正好前一子(在此範例中之子1)的範圍外面，及其次，在其名稱中之對應位元組必須小於或等於用於子2的範圍分界符。在此範例中，用於子2的範圍分界符係由值dcfa之2個位元組所描述。所以，用於候選元件之2個對應位元組必須小於或等於dcfa。使用此方法，在樹狀節點中之候選元件及所有的子可被檢驗，以檢查該候選元件將落在N+1個範圍中的那一個。針對第3F圖中所示之範例，若候選元件之名稱的4個對應位元組係大於其係f3231929之用於子N的鏈路之不同位元組的值時，則失誤情形將被偵測出。

【0143】樹狀導航處理可予以修正，而使此新的範圍節點適應。在到達範圍節點時，為了要選擇從該節點所發出之給定的鏈路，來自候選者之名稱的位元組必須落在被界定用於該特殊之鏈路的範圍內。若來自候選者之名稱的位元組之值大於在所有該等鏈路中的對應位元組之值時，則該候選元件落在由下面之子樹狀物所通過橫跨之所有範圍的外面---在此情況中(稱作“超出範圍情形”)，失誤情形被偵測出且樹狀導航處理終止。若來自候選元件之名稱的前導位元組落在由沿著通往子子樹狀物之鏈路之對應的不同位元組所決定的範圍內時，則樹狀導航處理繼續到下面的該子樹狀物。除非由於“超出範圍情形”而終止，否則樹狀導航可逐步向下地繼續到樹狀物之更深處，直至其到達葉狀節點資料結構為止。

【0144】此種範圍節點可與第3A至3E圖中所敘述之前綴樹節點結合，而在樹狀結構中被採用。在某些實施例中，樹狀結構的上方節點之一定數目的層次可係前綴樹節點，而樹狀遍歷則根據候選元件之名稱的前導位元組與沿著樹狀物之鏈路的對應位元組之間的精確匹配。隨後的節點可係範圍節點，而樹狀遍歷則由其中候選元件之名稱的前導位元組所落在之範圍來支配。在樹狀導航處理之中止時，如稍早在此檔案中所敘述地，可使用各種準則以決定應回報何者作為整體之內容相關查找的結果。

【0145】僅針對描繪和說明之目的，已提出用以表示及使用樹狀節點及葉狀節點之上述方法和設備的說明。然

而，它們並不打算要詳盡無遺的或限制本發明至所揭示的形式。從而，許多修正例和變化例將顯而易見於熟習本領域之從業者。

【0146】當候選元件被呈現作為輸入時，可遍歷上述之樹狀節點及葉狀節點結構，且可根據該候選元件之內容而執行樹狀物的內容相關查找。候選元件的名稱將由該候選元件之位元組所建構，正如當主要資料元件被安裝於節之中時，該主要資料元件的名稱係由該主要資料元件的內容所建構一樣。給定輸入之候選元件，用於樹狀物之內容相關查找的方法包含使用候選元件的名稱之樹狀結構的導航，其次藉由隨後的分析和篩查來決定應回報何者作為整體之內容相關查找的結果。換言之，樹狀導航處理回報第一輸出，且然後，分析和篩查被執行於該輸出上，而決定整體之內容相關查找的結果。

【0147】若有任何主要資料元件而其具備與候選者之名稱相同的前導位元組(或使得它們落在相同的範圍之內)時，則樹狀物將以由鏈路所表示之元件子集的形式而識別該子集之主要資料元件。一般而言，各樹狀節點或葉狀節點可儲存資訊，而使樹狀導航處理能根據輸入元件之名稱的對應位元組，而決定(若有的話)將選擇那一個外向鏈路以導航至樹狀物中之下一個較低的層次，及當該樹狀物係沿著所選擇的鏈路而被導航時所到達之節點的本體。假如各節點包含此資訊時，則樹狀導航處理可遞歸地向下導航至樹狀物中之各層次，直至沒有匹配被發現(在其中樹狀

導航處理可回報存在於被植根在目前節點處之子樹狀物中的一組主要資料元件之點)，或到達主要資料元件(在其中樹狀導航處理可回報主要資料元件及任何相關聯的元資料之點)為止。

【0148】一旦樹狀導航處理已終止，其它的準則和要求可被用來分析及篩查樹狀導航處理之輸出，而決定應回報何者作為整體之內容相關查找的結果。首先，可以用來自名稱之與候選元件共有之最多數量的前導位元組挑選主要資料元件。其次，當單一主要資料元件或多個主要資料元件係由樹狀導航處理所回報時，在給予將被回報作為該內容相關查找的結果之資格前，可具有額外的要求，亦即，它們應與候選元件的名稱共享某一最小數量的位元組(否則，該內容相關查找回報失敗)。篩查要求的另一範例可係其中，若該樹狀導航處理終止而未到達單一的主要資料元件，以致使多個主要資料元件(被植根於其中該樹狀導航終止之節點處)被回報作為樹狀導航處理的輸出時，則除非該等元件的數目小於諸如4至16個元件之某一特定的限制值，該等多個主要資料元件才將被給予回報作為該內容相關查找的結果之資格(否則，該內容相關查找回報失敗)。多個要求的組合可被採用以決定內容相關查找的結果。若多個候選元件仍保持時，則可檢驗導航預看位元組及相關聯的元資料，以決定那些主要資料元件係合適的。若仍無法將選擇縮小至單一的主要資料元件時，則可提供多個主要資料元件至衍生函數。以這種方式，查找處

理將報告“失敗”或回報單一主要資料元件，或者若非單一的主要資料元件時，則一組主要資料元件可能成為用以取得該候選元件之好的起點。

【0149】樹狀物需被設計用於有效率的內容相關存取。均衡的樹狀物將提供用於許多資料之存取的可比較深度。所期望的是，樹狀物之上方幾個層次將經常駐存於處理器快取中，下一幾個層次在快速記憶體中，以及隨後的層次在快閃儲存中。對於非常大的資料組，一或多個層次需駐存在快閃儲存中，且甚至在碟片中，是可能的。

【0150】第4圖顯示根據在此所敘述的一些實施例之256TB的主要資料可如何以樹狀形式組織之範例，並呈現該樹狀物可如何被佈局在記憶體及儲存中。假設每節點64個(其係 2^6 個)子的平均扇出，則用於主要資料元件之參照可藉由到達葉狀節點資料結構來存取(例如，如第3H圖中所敘述)，該葉狀節點資料結構係駐存在樹狀物的第6個層次(平均值)處(亦即，在5個鏈路遍歷或躍程之後)。所以，在5個躍程後之樹狀物的第6個層次處之該結構將駐存另外 2^{30} 個該節點在旁邊，各具有平均64個子(該等子係對主要資料元件的參照)，而藉以容納大約64個十億個主要資料元件。以4KB之元件大小計，此容納大約256TB個主要資料元件。

【0151】樹狀物可被佈局使得樹狀物的6個層次可被如下地遍歷：3個層次駐存在晶片上快取中(包含大約四千個“上方層次”之樹狀節點資料結構，而指明用於對大約

256K個節點之鏈路的轉換)，2個層次在記憶體中(包含16個百萬個“中間層次”之樹狀節點資料結構，而指明用於對大約十億個葉狀節點之鏈路的轉換)，以及第6個層次在快閃儲存中(容納十億個葉狀節點資料結構)。駐存在快閃儲存中之樹狀物的此第6個層次處之該十億個葉狀節點資料結構提供用於該64個十億個主要資料元件的參照(每葉狀節點平均64個元件)。

【0152】在第4圖中所示的範例中，在第4及第5個層次處，各節點致力於每元件平均16個位元組(例如，1個位元組用於子ID，6個位元組用於對PDE的參照，加上1個位元組用於位元組計數，加上平均8個位元組用以指明實際的轉換位元組以及一些元資料)。在第6個層次處，各節點致力於每元件平均48個位元組(1個位元組用於子ID，1個位元組用於位元組計數，8個位元組用以指明實際的轉換位元組，6個位元組用於對主要資料元件的參照，1個位元組用於離此主要資料元件之衍生物的計數，16個位元組用於導航預看，2個位元組用於主要資料元件的大小，以及13個位元組之其它的元資料)，因此，在用於樹狀物(包含對主要資料元件的參照，及包含任何元資料)所需之快閃儲存中的總容量係大約3萬億個位元組。用於樹狀物之上方節點所需的總容量係此規模的小部分(因為有較少的節點，且需較少的位元組用以指明對子節點之較緊的參照，以及每個節點需要較少的元資料)。在該範例中，該等上方樹狀節點致力於每元件平均8個位元組(1個位元組用於

子ID，1個位元組用於位元組計數，加上平均3至4個位元組用以指明實際的轉換位元組，以及2至3個位元組用於對子節點的參照)。總體而言，在此範例中，具有256TB之主要資料的合成資料組係使用3TB(或256TB的1.17%)之額外的設備來排序成一10億個群組。

【0153】在第4圖中所示的範例中，當256TB的主要資料包含64個十億個各4KB之主要資料元件時，需要少於5個位元組(或36個位元)的位址以完全區分該64個十億個主要資料元件。從內容相關之立場，若資料的混合係使得進行之名稱的平均4個位元組被消耗在最初3個層次之各者處，以及8個位元組在其次3個層次之各者處，則總計36個位元組(288個位元)之名稱(平均值)將區分所有64個十億個主要資料元件。該等36個位元組將比構成各元件之4KB的1%更小。若4KB之主要資料元件可由其位元組的1%(或甚至5至10%)所識別時，則剩餘部分的位元組(其構成多數的位元組)可忍受擾動，且具有該等擾動的候選元件仍可到達此主要資料元件，並可針對來自它的衍生來考慮。

【0154】應注意的是，在任一給定鏈路上之位元組的數目(用以區分底下之各種子樹狀物)，將由包含資料組之混合元件中的實際資料所支配。同樣地，在給定節點當中之鏈路的數目亦將隨著該資料而變化。自描述樹狀節點及葉狀節點資料結構將宣告用於各鏈路所需之位元組的實際數目和值，以及從任一節點所發出之鏈路的數目。

【0155】進一步的控制可被安置用以限制所致力於樹

狀物的各種層次處之快取、記憶體及儲存的數量，而在增量之儲存的預算分配內盡可能地將輸入排序成許多不同的群組。為了要操縱其中有資料之密度和封包而其需要很深的子樹狀物以充分區分元件之情勢，該等密度可藉由群組較大組之相關的元件成為樹狀物之某一深度(例如，第6個層次)的平面群組，且在該等者之上執行精簡的搜尋和衍生(藉由先檢驗導航預看及元資料以決定最佳的主要資料元件，否則(作為備用)僅針對剩餘部分之資料搜尋複製品，而非由該方法所給予之充分的衍生)來有效率地操縱。此將規避很深之樹狀物的建立。另一替代例則是只要該等層次適合可用的記憶體，就允許深的樹狀物(具有許多層次)。在較深之層次溢出至快閃或碟片時，可採取步驟而從該層次起將樹狀物平坦化，用以使得將在其它方面由於對儲存在快閃或碟片中之更深層次樹狀節點的多個連續存取而被引致之等待時間最小化。

【0156】所期望的是，來自元件之名稱的合計位元組之相對小的部分將經常足以識別各主要資料元件。使用在此所敘述的實施例而在各種真實世界資料集上所執行之研究證實的是，主要資料元件的位元組之小的子集可用以排列多數的元件，而致能解決方法。因此，就用於其操作所需求之儲存的數量而言，這種解決方法係有效率的。

【0157】就來自第4圖之範例所需的存取而言，一旦用於輸入(或候選元件)之每個進來的4KB資料塊，該方案將需要以下之存取以查詢樹狀結構及到達葉狀節點：三個

快取參照、兩個記憶體參照(或多個記憶體參照)，加上來自快閃儲存之單一IO以存取葉狀節點資料結構。來自儲存之此單一IO將提取4KB頁面，其將保持用於大約64個元件的群組之葉狀節點資料結構的資訊，該資訊將包含48個位元組以致力於所討論的主要資料元件。該等48個位元組將包含所討論之主要資料元件上的元資料。此將得出樹狀查找處理的結論。之後，所需之IO的數目將根據候選元件是否變成複製品、衍生物或將被安裝在篩中之新的主要資料元件而定。

【0158】其係主要資料元件之複製品的候選元件將需要一個IO用以提取主要資料元件，以便驗證該複製品。一旦該複製品被驗證，就將有再一個IO用以更新樹狀物中的元資料。因此，在該樹狀查找之後，複製品元件的攝取將需要兩個IO，而總計3個IO。

【0159】樹狀查找失敗的，以及不是複製品亦非衍生物的候選元件需要再一個IO用以儲存該元件於篩中作為新的主要資料元件，和另一個IO用以更新樹狀物中的元資料。因此，在樹狀查找之後，該樹狀查找失敗之候選元件的攝取將需要兩個IO，而導致總計3個IO。然而，對於其中樹狀查找處理終止而無需儲存IO的候選元件，僅需總計2個IO以供攝取該等候選元件之用。

【0160】衍生物(但非複製品)的候選元件將首先需要一個IO以提取用以計算衍生所需之主要資料元件。因為所期望的是，最經常的衍生將離開單一的主要資料元件(而

非多個)，所以僅需要一個IO以提取主要資料元件。在成功完成衍生之後，將需要再一個IO用以在用於儲存中之元件所建立的條目中儲存重建程式及衍生細節，和另一個IO用以更新樹狀物中的元資料(諸如計數，等等)而反映新的衍生物。因此，在該首先樹狀查找之後，變成衍生物之候選元件的攝取需要3個額外的IO，而總計4個IO。

【0161】綜上所述，用以攝取候選元件及施加資料提取(Data Distillation™)方法至其(同時橫跨很大的資料集全面地利用冗餘)，需要大約3至4個IO。與傳統之重複資料刪除技術所需要的相比，此一般僅係每候選元件多一個IO，作為回報的是，冗餘可以用比元件本身更精細的粒度而橫跨資料集被全面地利用。

【0162】每秒提供250,000個隨機IO存取(其意指對4KB頁面之1GB/秒的隨機存取頻寬)之儲存系統可在大約每秒62,500個輸入資料塊(由各平均大小4KB之每輸入資料塊4IO除250,000)上攝取及執行資料提取(Data Distillation™)方法。此在用盡該儲存系統之全部頻寬的同時，致能250MB/秒的攝取速率。若僅使用儲存系統的一半頻寬時(使得另一半可用於對所儲存資料之存取)，該重複資料刪除系統仍可傳遞125MB/秒的攝取速率。因而，當給定足夠的處理功率時，資料提取(Data Distillation™)系統能在現代儲存系統上以經濟的IO全面地利用冗餘橫跨資料集(以比元件本身更精細的粒度)，且以每秒數百個百萬位元組的攝取速率傳遞資料縮減。

【0163】因此，如測試結果所證實的，在此所敘述之實施例以經濟的IO存取及以用於設備所需的最小增量儲存，達成自大規模儲存之資料而搜尋元件之複雜任務(其中輸入元件可以用指明衍生所需的最小儲存取得)。因而所建構的此框架可使用元件之合計位元組的較小百分比以發現適合於衍生之元件，而留下可用於擾動及衍生之大量的位元組。解釋此方案為何有效率地作功於許多資料的重要見解在於，樹狀物提供易於使用的、精細粒度的結構，而允許設置可識別篩中元件之不同的和有區別的位元組，且雖然該等位元組係各自在資料中之不同的深度及位置處，但它們可在樹狀結構中被有效率地隔離和儲存。

【0164】第5A至5C圖顯示資料可如何使用在此所敘述之實施例而被組織的實際範例。第5A圖顯示512個位元組的輸入資料，及分解的結果(例如，在第2圖中之執行操作202的結果)。在此範例中，指紋圖譜係施加以決定資料中的斷點，以致使連續的斷點識別候選元件。交替的候選元件已使用粗體及常規字體予以顯示。例如，第一候選元件係“b8ac83d9dc7caf18f2f2e3f783a0ec69774bb50bbe1d3ef1ef8a82436ec43283bc1c0f6a82e19c224b22f9b2”，以及下一個候選元件係“ac83d9619ae5571ad2bbcc15d3e493eef62054b05b2dbccce933483a6d3daab3cb19567dedbe33e952a966c49f3297191cf22aa31b98b9dcd0fb54a7f761415e”，等等。第5A圖的輸入係分解成如所示之12個可變大小的候選元件。各資料塊的前導位元組係使用來在篩中排列及組織元件。

第5B圖顯示第5A圖中所示之12個候選元件可如何使用它們的名稱，及使用第3B圖中所敘述之樹狀結構，而在節中被以樹狀物組織成為主要資料元件。各元件具有不同的名稱，其係由元件的整個內容所建構。在此範例中，因為指紋圖譜係施加以決定該12個候選元件之間的斷點，所以各候選元件的前導位元組將已經與定錨指紋圖譜對齊；因此，各名稱的前導位元組將已由被定錨在此指紋圖譜處之內容的第一維度所建構。該名稱的該等前導位元組組織各種元件。例如，若在元件的名稱中之第一位元組係等於“0x22”時，則取頂部鏈路而選擇主要資料元件#1。應注意的是，第5B圖中之各種鏈路係使用不同數目的位元組來區分，如參照第3B圖中所描繪之樹狀資料結構所解說的。

【0165】第5C圖顯示第5A圖中所示之12個候選元件可如何使用參照第3D圖所敘述的樹狀資料結構來組織。指紋圖譜係進一步施加至各元件的內容，用以識別元件內容之內的次要指紋圖譜。自第一指紋圖譜(已經存在於各元件的邊界處)及第二指紋圖譜之位置所提取之內容的位元組係串聯以形成名稱的前導位元組，其係使用來組織該等元件。換言之，元件名稱係建構如下：來自兩個維度或欄位之資料的位元組(由定錨指紋圖譜及次要指紋圖譜所定位)係串聯以形成名稱的前導位元組，其次係剩餘部分之位元組。作為該名稱結構之此選擇的結果，不同序列之位元組導致第5C圖中的各種主要資料元件(與第5B圖相對地)。例如，用以到達主要資料元件#4，樹狀導航處理首

先取用對應“46093f9d”(其係在第一維度(亦即，第一指紋圖譜)處之欄位的前導位元組)之鏈路，且然後，取用對應“c4”(其係在第二維度(亦即，第二指紋圖譜)處之欄位的前導位元組)之鏈路。

【0166】第6A至6C圖顯示根據在此所敘述的一些實施例之樹狀資料結構可如何使用於參照第1A至1C圖所分別敘述的內容相關映射器121及122。

【0167】一旦已解決搜索合適之主要資料元件(企圖要從它來取得候選元件)的困難問題，問題就會縮小到檢驗主要資料元件之一者或小的子集，及以用以指名衍生所需之最小儲存自它們最佳地取得候選元件。其它目標包含保持對儲存系統之存取的數目至最小，以及保持衍生時間及重建時間可接受。

【0168】衍生器必須表示候選元件為被執行在一或多個主要資料元件上之轉變的結果，且必須指明該等轉變為將被用來在資料檢索時再生衍生物的重建程式。各衍生可要求它自己唯一的程式將被建構。衍生器的功能在於識別該等轉變，且以最小的足跡建立重建程式。各種轉變可被採用，包含被執行於一或多個主要元件上或各元件的特定欄位之上的算術、代數或邏輯運算。此外，可在一或多個主要資料元件中使用諸如，位元組的串聯、插入、置換及刪除的位元組處理轉變。

【0169】第7A圖提供根據在此所敘述的一些實施例之可在重建程式中被指明的轉變之範例。在此範例中所指明

的詞彙包含在元件中的指明長度之欄位上的算數運算，以及在主要資料元件中的指明偏移量處之位元組宣告長度的插入、刪除、附加及置換。各種技術及運算可由衍生器所採用，用以偵測候選元件與主要資料元件之間的相似度及差異，以及用以建構重建程式。衍生器可利用可用於基本硬體中的詞彙以執行它的功能。工作的最終結果在於指明被指明用於重建程式之詞彙中的轉變，且在於使用最小數量的增量儲存及以亦致能快速資料檢索的方式如此進行。

【0170】衍生器可利用基本機器的處理能力，且在所分配至它的處理預算內工作，而可在系統的成本-效能約束內提供最佳的分析可能。鑑於微處理器核心更容易獲得，且鑑於對儲存之IO存取昂貴，資料提取(Data Distillation™)解決方法已被設計要利用現代微處理器的處理能力，而在少許的主要資料元件上有效率地執行候選元件之內容的局部分析和衍生。所期望的是，(對於非常大的資料)資料提取(Data Distillation™)解決方法的效能將不受計算處理的速率限制，而是受到典型儲存系統的IO頻寬的速率限制。例如，所期望的是，兩個微處理器核心就將足以在支援250,000個IO/秒之典型以快閃為主的儲存系統上，執行用以支援每秒數百個百萬位元組之攝取速率的所需計算和分析。應注意的是，來自諸如Intel Xeon處理器E5-2687W(10核心，3.1GHz，25MB快取)之現代微處理器的兩個該微處理器核心，係可獲自處理器之總計算功率的一小部分(十分之二)。

【0171】第7B圖顯示根據在此所敘述的一些實施例之將從主要資料元件得到的候選元件之結果的範例。具體地，資料圖案“Elem”係主要資料元件，其係儲存在主要資料篩中，以及資料圖案“Cand”係候選元件，其將衍生自該主要資料元件。在“Cand”與“Elem”之間的18個共同的位元組已被凸顯。重建程式702指明資料圖案“Cand”可如何衍生自資料圖案“Elem”。如第7B圖中所示，重建程式702描繪如何藉由使用1個位元組置換、6個位元組插入、3個位元組刪除、7個位元組大批置換，而自“Elem”取得“Cand”。用以指明衍生物之成本係20個位元組+3個位元組的參照=23個位元組，其係原始大小的65.71%。應注意的是，所顯示的重建程式702係人類可讀取表示之程式，且並非如何藉由在此所敘述的實施例來實際儲存之程式。同樣地，根據諸如乘法及加法之算術運算的其它重建程式亦已在第7B圖中被顯示。例如，若“Elem”係bc1c0f6a790c82e19c224b22f900ac83d9619ae5571ad2bbec152054ffffff83，以及“Cand”係bc1c0f6a790c82e19c224b22f91c4da1aa0369a0461ad2bbec152054ffffff83，則8個位元組差異可使用相乘 $(00ac83d9619ae557) * 2a = [00]1c4da1aa0369a046$ 來如所示地取得。用以指明衍生物之成本：4個位元組+3個位元組的參照=7個位元組，其係原始大小的20.00%。選擇性地，若“Elem”係bc1c0f6a790c82e19c224b22f9b2ac83ffffffffffffffffffffffff283，以及“Cand”係bc1c0f6a790c82e19c224b22f9b2ac83000000000000000000000000000000002426，則

16個位元組差異可使用加法而被如所示地取得，例如，藉由將0x71a3相加至開始於偏移16之16個位元組區域，且拋棄進位。用以指明衍生物之成本係5個位元組+3個位元組的參照=8個位元組，其係原始大小的22.85%。應注意的是，在第7A圖中之該等取樣編碼僅係選擇用於描繪之目的。在第7B圖中之範例具有32個位元組的資料大小，且因此，5個位元滿足元件內之長度及偏移欄位。對於大的元件(例如，4KB元件)，該等欄位的大小將需要被增加至12個位元。同樣地，取樣編碼容納3個位元組或24個位元的參照大小。此應允許16個百萬個主要資料元件將被參照。若參照必須能在例如，256TB之資料中定址任何位置，則該參照在大小上將必須變成6個位元組。當該資料集被分解成4KB元件時，用以指明參照所需的6個位元組將係4KB元件之大小的一小部分。

【0172】用以指明衍生物元件(其係衍生自一或多個主要資料元件)所需之資訊的大小係重建程式之大小及用以指明所需之(一或多個)主要資料元件所需的參照之大小的總和。用以指明候選元件作為衍生物元件所需之資訊的大小係稱作候選元件距離主要資料元件的距離。當候選元件可切實衍生自多組主要資料元件之任時，則具有最短距離的一組主要資料元件將被選定作為目標。

【0173】當候選元件必須被衍生自超過一個的主要資料元件時(藉由組裝所衍生自該等者之各者的提取)，則衍生器必須考慮對儲存系統的額外存取之成本中的因子，以

及對較小重建程式和較小距離之利益的衡量。一旦最佳重建程式已被建立用於候選者，就將其距離與距離臨限值相比；若不超過該臨限值時，則接受衍生。一旦接受該衍生，該候選元件就被重制定作為衍生物元件，且由主要資料元件及重建程式的組合所置換。在被建立用於該候選元件的提取資料中的條目，係由重建程式加上對相關聯之主要資料元件的一或多個參照所置換。若用於最佳衍生之距離超過該距離臨限值時，則該衍生物將不被接受。

【0174】為了要產生資料縮減，距離臨限值必須一直小於候選元件的大小。例如，距離臨限值可被設定為候選元件之大小的50%，使得除非衍生的足跡係小於或等於候選元件之足跡的一半，該衍生才將被接受，而藉以確保兩倍或更大的縮減用於其中存在合適衍生的各候選元件。距離臨限值可係根據使用者特定輸入，或由系統所選定之預定的百分比或分數。該距離臨限值可根據系統之靜態或動態的參數，而由系統所決定。

【0175】第8A至8E圖顯示根據在此所敘述的一些實施例之資料縮減可如何藉由將輸入資料分解成固定大小元件，且以參照第3D及3E圖所敘述的樹狀資料結構組織該等元件來執行。第8A圖顯示輸入資料如何可被簡單地分解成32個位元組資料塊。具體地，第8A圖顯示最初的10個資料塊，且然後，再多些的資料塊，其稍後呈現例如，42個百萬個資料塊。第8B圖顯示篩中之主要資料元件的組織，其係使用所建構的名稱以致使該名稱的前導位元組包含來

自元件內容中之3個維度的內容(對應定錨指紋圖譜、次要指紋圖譜及第三指紋圖譜的位置)。具體地，在第8B圖中，各32個位元組資料塊變成32個位元組(固定大小區塊)的候選元件。指紋圖譜的方法係施加至元件的內容。各元件具有名稱，其係建構如下：來自元件的三個維度或欄位(分別由定錨指紋圖譜、次要指紋圖譜及第三指紋圖譜所定位)之資料的位元組係串聯以形成名稱的前導位元組，其次係該元件之剩餘部分的位元組。該名稱係使用來組織篩中的元件。如第8B圖中所示，最初的10個資料塊不包含複製品或衍生物，且係接連地安裝作為篩中的元件。第8B圖顯示在消耗第10個資料塊之後的篩。第8C圖顯示在消耗資料輸入的額外若干個百萬個元件之後，例如，在呈現接著的42個百萬個資料塊之後的時間中之隨後點處的篩之內容。該篩係檢驗用於複製品或衍生物。無法衍生自元件的資料塊被安裝在篩中。第8C圖顯示在消耗42個百萬個資料塊之後的篩，包含例如，16,000,010個元件(可以用3個參照位址邏輯地定址)，而剩餘的26,000,000個資料塊變成衍生物。第8D圖顯示新的輸入之範例，其係隨後提交至篩且被識別為篩中之條目(被顯示為元件號碼24,789)的複製品。在此範例中，篩識別元件24,789(資料塊9)作為用於資料塊42,000,011之最合適的元件。提取功能決定的是，新的資料塊係精確的複製品，且以對元件24,789的參照置換它。用以表示該衍生物的成本係與35B原始相對的3個位元組，其係原始大小的8.57%。第8D圖顯示輸入的第二個範

例(資料塊 42,000,012)，其係轉換成篩中之條目(被顯示為元件號碼 187,126)的衍生物。在此範例中，篩決定的是，並沒有精確的匹配。其識別元件 187,125 及 187,126(資料塊 8 及 1)作為最合適的元件。新的元件係衍生自最合適的元件。衍生對元件 187,125 及衍生對元件 187,126 係描繪於第 8D 圖中。用以表示衍生對元件 187,125 之成本係 39 個位元組 + 3 個位元組的參照 = 42 個位元組，其係原始大小的 120.00%。用以表示衍生對元件 187,126 之成本係：12 個位元組 + 3 個位元組的參照 = 15 個位元組，其係原始大小的 42.85%。最佳的衍生(對元件 187,126)被選定。重建大小係與臨限值相比。例如，若臨限值係 50%，則此衍生物(42.85%)被接受。第 8E 圖提供從主要資料元件所取得的資料塊之兩個額外的範例，包含衍生物係藉由衍生自兩個主要資料元件來實際地建立的一範例。在第一範例中，資料塊 42,000,013 被提出。篩識別元件 9,299,998(資料塊 10)作為最合適的元件。衍生對元件 9,299,998 係描繪於第 8E 圖中。用以表示該衍生物之成本係 4 個位元組 + 3 個位元組的參照 = 7 個位元組，其係原始大小的 20.00%。重建大小係與臨限值相比。例如，若臨限值係 50%，則此衍生物(20.00%)被接受。在第二範例中，資料塊 42,000,014 被提出。在此範例中，資料塊 42,000,014 係使得資料塊的一半可最佳地衍生自元件 9,299,997，而資料塊的另一半可最佳地衍生自元件 9,299,998。因此，多個元件之衍生物被建立出，而產生進一步的資料縮減。多個元件之衍生物係顯示於

第8E圖中。用以表示此多個元件之衍生物的成本係3個位元組的參照+3個位元組+3個位元組的參照=9個位元組，其係原始大小的25.71%。重建大小係與臨限值相比，例如，若臨限值係50%，則此衍生物(25.71%)被接受。應注意的是，來自單一元件衍生物的最佳輸出將係45.71%。

【0176】第8A至8E圖顯示資料提取(Data Distillation™)系統的重要優點：可在執行資料縮減之同時，有效消耗及產生固定大小的區塊。應注意的是，固定大小的區塊係在高效能儲存系統中高度期望的。使用資料提取(Data Distillation™)設備，包含許多固定大小區塊之大的輸入之輸入檔案可被分解成許多固定大小的元件，以致使所有的主要資料元件是固定大小的。用於各衍生物元件之潛在可變大小的重建程式可被包裝在一起，且被保持內嵌在提取的資料檔案中，而可被隨後組塊成為固定大小的區塊。因此，針對所有實用之目的，可執行強大的資料縮減，且同時在儲存系統中消耗及產生固定大小的區塊。

【0177】第9A至9C圖顯示其係最初顯示於第1C圖中之資料提取(Data Distillation™)方案的範例：此方案採用個別的主要重建程式篩，其可以用內容相關方式存取。這種結構致能衍生物的偵測，其建構已經存在於主要重建程式篩中的重建程式。這種衍生物可被重制定以參照現有的重建程式。此致能了重建程式中之冗餘的偵測。在第9A圖中，輸入資料被攝取。指紋圖譜的方法係施加至該資料，以及資料塊邊界係設定在指紋圖譜位置處。該輸入係分解

成如所顯示之8個候選元件(交變的資料塊以粗體及常規字體顯示於第9A圖中)。在第9B圖中，該8個候選元件係如在篩中所組織地顯示。各元件具有不同的名稱，其係由該元件的整個內容所建構。在此範例中，元件名稱係建構如下：來自兩個維度或欄位之資料的位元組(分別由定錨指紋圖譜及次要指紋圖譜所定位)係串聯以形成名稱的前導位元組，其次係剩餘部分之位元組。該名稱係使用來在篩中排列元件，且亦透過樹狀結構而提供內容相關存取。第9B圖亦提供第二內容相關結構，其包含主要重建程式。第9C圖顯示複製品重建。假定並非任一主要資料元件之複製品的55個位元組候選元件(第9C圖中所示)到達。元件3係選擇作為最合適的元件---對於PDE 2及3，最初的兩個維度係相同的，但以88a7開始之剩餘部分的位元組匹配元件3。新的輸入係以12個位元組重建程式(RP)衍生自元件3。編碼係如第7A圖中所示。應注意的是，用於此範例，最大元件大小係64位元，以及例如，與第7A圖中所示之5位元的長度和偏移相對地，所有的偏移和長度係編碼為6位元的值。該RP儲存被搜尋，且此新的RP並未被發現。此RP係插入至主要RP儲存內，而根據其值予以排列。新的元件係重制定為對主要資料元件3的參照，及對在RP儲存中之參照4處的新建立之主要重建程式的參照。用於此取得之元件的總儲存大小係：3個位元組的PDE參照、3個位元組的RP參照、12個位元組的RP=18個位元組，其係與儲存它為PDE相對之大小的31.0%。之後，假定55個位元組候選

元件的副本到達。如之前一樣地，12個位元組RP係根據元件3而被建立出。RP儲存被搜尋，且具有主要RP ID=3，RP參照=4的RP被發現。此候選元件係表示於系統中作為對主要資料元件3的參照，及對重建程式4的參照。所添加用於此取得之元件的總儲存大小現在係：3個位元組的PDE參照、3個位元組的RP參照=6個位元組，其係與儲存它為PDE相對之大小的10.3%。

【0178】第10A圖提供根據在此所敘述的一些實施例之在重建程式中所指明的轉變如何施加至主要資料元件以產生衍生物元件之範例。該範例顯示被指明將由編號187,126之主要資料元件(此主要資料元件亦被顯示於第8C圖中的篩之中)所產生(藉由施加如所示之重建程式所指明的四種轉變(插入、置換、刪除及附加)至它)的衍生物元件。如第10A圖中所示，元件187,126係自篩載入，且重建程式被執行以自元件187,126取得資料塊42,000,012。第10B至10C圖顯示根據在此所敘述的一些實施例之資料檢索處理。各資料檢索請求本質地採取所提取的資料中之元件的形式，以無損縮減形式提交至搜尋引擎。用於各元件的無損縮減形式包含對相關聯之主要資料元件及重建程式的參照。資料提取(Data Distillation™)設備的檢索器提取主要資料元件及重建程式，且提供該等者至用於重建的重建器。在已提取用於所提取的資料之元件的有關聯主要資料元件及重建程式之後，重建器執行重建程式，而以元件之原始未縮減形式產生元件。資料檢索處理所需之用以執

行重建的努力係線性相對於重建程式的大小和主要資料元件的大小。因此，高的資料檢索速率可藉由系統來達成。

【0179】顯而易見的是，用以將元件自提取資料中之無損縮減形式重建至其原始未縮減形式，僅需提取被指明用於該元件的主要資料元件及重建程式。因此，為了要重建給定的元件，並不需要存取或重建其它的元件。這使得該資料提取(Data Distillation™)設備有效率，即使當服務用於重建及檢索之隨機的請求序列時。應注意的是，諸如Lempel-Ziv方法之傳統的壓縮方法必須將包含期望區塊之整個窗口的資料提取及解壓縮。例如，若儲存系統採用Lempel-Ziv方法而使用32KB之窗口來壓縮4KB區塊的資料時，為了要提取及解壓縮給定的4KB區塊，則必須將32KB的整個窗口擷取及解壓縮。這會帶來效能上的損失，因為為了要傳遞期望之資料，更多的頻寬會被消耗，且更多的資料需被解壓縮。所述資料提取(Data Distillation™)設備並不會招致該損失。

【0180】資料提取(Data Distillation™)設備可以用各種方式被整合至電腦系統內，而可在系統中全域地跨整個資料的範圍有效率地揭露和利用冗餘之方式來組織及儲存資料。第11A至11G圖顯示根據在此所敘述的一些實施例之包含資料提取(Data Distillation™)機制(其可使用軟體、硬體或其組合來實施)的系統。第11A圖呈現具有軟體應用的通用型計算平台，該軟體應用運行於系統軟體上，該系統軟體執行於包含處理器、記憶體及資料儲存組件的硬體

平台上。第11B圖顯示被整合至平台之應用層內的資料提取(Data Distillation™)設備，而各特定之應用使用該設備來利用用於該應用的資料集內之冗餘。第11C圖顯示被採用來提供資料虛擬層或服務於被運行於其上之所有應用的資料提取(Data Distillation™)設備。第11D及11E圖顯示具有取樣計算平台的作業系統、檔案系統及資料管理服務之資料提取(Data Distillation™)設備的兩種不同形式的整合。其它的整合方法包含(但並未受限於)與在如第11F圖中所示之諸如，被採用於以快閃為主之資料儲存子系統中之硬體平台中的嵌入式計算堆疊之整合。

【0181】第11G圖呈現與第11D圖中所示之取樣計算平台的資料提取(Data Distillation™)設備之整合的額外細節。第11G圖顯示資料提取(Data Distillation™)設備的組件，其具有在通用型處理器上被執行為軟體的解析器和分解器、衍生器、檢索器及重建器，以及駐存在儲存階層的幾個層次之範圍的內容相關映射結構。主要資料節可駐存在儲存媒體(諸如以快閃為主的儲存裝置)中。

【0182】第11H圖顯示資料提取(Data Distillation™)設備可如何與取樣通用型計算平台介接。

【0183】檔案系統(或filesystem)使檔案(例如，文本文件、電子表格、可執行的、多媒體的檔案、等等)與識別符(例如，檔名、檔案處置代碼、等等)相關聯，且使操作(例如，讀取、寫入、插入、附加、刪除、等等)能藉由使用與檔案相關聯的識別符而在檔案上被執行。由檔案系

統所實施的命名空間可以是平面的或階層的。此外，命名空間可被分層，例如，頂部層之識別符可在接連的下方層被解析成為一或多個識別符，直至該頂部層之識別符被完全地解析為止。以這種方式，檔案系統提供實體地儲存檔案的內容之實體儲存裝置及/或儲存媒體(例如，電腦記憶體、快閃驅動器、碟片驅動器、網路儲存裝置、CD-ROM、DVD、等等)的抽象概念。

【0184】用以在檔案系統中儲存資訊所使用的實體儲存裝置及/或儲存媒體可使用一或多個儲存技術，且可被設置在相同的網路位置處或可跨不同網路位置來分配。鑒於與檔案相關聯的識別符及被請求要在該檔案上執行的一或多個操作，檔案系統可(1)識別一或多個實體儲存裝置及/或儲存媒體，以及(2)致使由該檔案系統所識別的該等實體儲存裝置及/或儲存媒體，實行所被請求要在與該識別符相關聯之檔案上執行的操作。

【0185】無論何時只要讀取或寫入操作被執行於系統之中，就可涉及不同的軟體及/或硬體組件。“讀取器”之用語可意指當給定的讀取操作被執行於系統之中時涉及之該系統中的軟體及/或硬體組件的集合，其是以及“寫入器”之用語可意指當給定的寫入操作被執行於系統之中時涉及之該系統中的軟體及/或硬體組件的集合。在此所敘述之用於資料縮減的方法及設備之一些實施例可藉由當執行給定的讀取或寫入操作時所涉及之系統的一或多個軟體及/或硬體組件來使用，或可被結合至當執行給定的讀取

或寫入操作時所涉及之系統之一或多個軟體及/或硬體組件內。不同的讀取器和寫入器可使用或結合不同的資料縮減實施。然而，使用或結合特殊之資料縮減實施的各寫入器將對應亦使用或結合相同之資料縮減實施的讀取器。應注意的是，在系統中所執行的某些讀取和寫入操作可不一定要使用或結合資料縮減設備。例如，當資料提取(Data Distillation™)設備或資料縮減設備103檢索主要資料元件，或添加新的主要資料元件至主要資料篩時，其可直接執行讀取和寫入操作，而無需資料縮減。

【0186】具體地，在第11H圖中，寫入器150W可一般意指當執行給定的寫入操作時所涉及的系統之軟體及/或硬體組件，以及讀取器150R可一般意指當執行給定的讀取操作時所涉及的系統之軟體及/或硬體組件。如第11H圖中所示，寫入器150W提供輸入資料至資料提取(Data Distillation™)設備或資料縮減設備103，且自資料提取(Data Distillation™)設備或資料縮減設備103接收提取之資料108。讀取器150R對於資料提取(Data Distillation™)設備或資料縮減設備103提供檢索請求109，且自資料提取(Data Distillation™)設備或資料縮減設備103接收檢索的資料輸出113。

【0187】用於第11H圖的實施範例包含(但不受限於)結合或使用資料提取(Data Distillation™)設備或資料縮減設備103於應用程式、作業系統核心、檔案系統、資料管理模組、裝置驅動器或快閃或碟片驅動器的韌體中。此橫

跨第11B至11F圖中所敘述之各種組態和使用。

【0188】第11I圖顯示如何將資料提取設備用於在區塊處理儲存系統中的資料縮減。在這種區塊處理系統中，資料被儲存在區塊中，並且每個區塊係由邏輯區塊位址或LBA來識別。區塊被不斷地修改和覆寫，使得新資料可以被覆寫到由特定LBA識別的區塊中。系統中的每個區塊被視為候選元件，並且資料提取設備可以用於將候選元件縮減為無損縮減形式，其包括對於(儲存在特定主要資料元件區塊中的)主要資料元件的參照，以及在衍生物元件的情況下，對於(儲存在特定的重建程式區塊中的)重建程式的參照。第11I圖介紹資料結構1151，其將由LBA識別的區塊的內容映射到無損縮減形式的對應元件。對於每個LBA將駐存相關元件的規範。對於採用固定大小的區塊的系統，使輸入區塊、主要資料元件區塊1152以及重建程式區塊1153全部都具有固定大小是方便的。在此系統中，每個主要資料元件可以儲存為單一區塊。多個重建程式可能被封裝到相同固定大小的重建程式區塊中。資料結構還包含對於計數(Count)欄位和駐存在每個主要資料元件和重建程式的葉節點資料結構處的關聯元資料的參照，以使當區塊被新資料覆寫時，駐存在LBA的先前資料可以被有效地管理--(正在被覆寫的)現有的主要資料元件和重建程式的計數欄位必須被遞減，同樣地，由輸入到LBA的資料所參照的主要資料元件的計數必須被遞增。藉由保持對此資料結構1151中的計數欄位的參照，可以快速地管理覆寫，

從而能夠充分利用由資料提取設備所提供的資料縮減的高效能區塊處理儲存系統。

【0189】第12A圖顯示根據在此所敘述的一些實施例之用於橫跨頻寬約束通訊媒體的資料通訊之資料提取(Data Distillation™)設備的使用。在所示之設置中，通訊節點A建立出將被傳送至通訊節點B的一組檔案。節點A採用資料提取(Data Distillation™)設備以將該等輸入檔案轉變成提取的資料或提取檔案，其包含對被安裝在主要資料篩中之主要資料元件的參照，以及用於衍生物元件的重建程式。節點A接著連同主要資料篩將提取檔案傳送至節點B(主要資料篩可在傳送提取檔案之前、同時或之後被傳送；此外，主要資料篩可在與被用來傳送提取檔案之通訊頻道相同的通訊頻道上，或不同的通訊頻道上被傳送)。節點B安裝主要資料篩於其端點之對應結構中，且隨後，透過駐存在節點B之資料提取(Data Distillation™)設備中的檢索器和重建器而饋送提取檔案，用以產生由節點A所建立出之原始組的檔案。因此，更有效率的使用是藉由採用資料提取(Data Distillation™)設備於頻寬約束通訊媒體的兩端點處僅傳送縮減之資料，而由該頻寬約束通訊媒體所做成。應注意的是，使用資料提取(Data Distillation™)將致能橫跨較大範圍之冗餘的利用(超過使用諸如Lempel-Ziv之習知技術所可看到的)，以致使即使是大型的檔案或成群組的檔案均可被有效率地傳輸。

【0190】現在我們討論資料提取(Data Distillation™)

設備在寬域網路設施中之使用，其中工作小組協作地共享被橫跨多個節點傳播之資料。當資料被最初地建立時，其可被如第12A圖中所描繪地縮減和傳達。寬域網路在各地點保持資料的副本，用以致能對該資料的快速本地存取。資料提取(Data Distillation™)設備的使用可縮減各地點的足跡。再者，在任一地點之新資料的隨後攝取時，在該新資料與現有的主要資料篩的內容間的任何冗餘可被利用以縮減該新的資料。

【0191】在這種設施中，對任何給定地點處之資料的任何修正必須被傳達至所有其它的地點，使得在各地點處之主要資料篩保持一致。因此，如第12B圖中所示，諸如主要資料元件之安裝及刪除的更新，以及元資料更新，可根據在此所敘述的一些實施例而被傳達至各地點處之主要資料篩。例如，在將新的主要資料元件安裝至給定地點處的篩之內時，該主要資料元件必須被傳達至所有其它的地點。各地點可使用該主要資料元件的值，而以內容相關方式來存取該篩，以及決定新的條目需被添加在該篩的何處。同樣地，在從給定地點處的篩刪除主要資料元件時，所有其它的地點必須被更新以反映該刪除。此可被完成的一方式係藉由將主要資料元件傳達至所有的地點，使得各地點可使用該主要資料元件來內容相關地存取篩，以及決定葉狀節點中的那一個條目需被刪除，連同對樹狀物中之相關鏈路的必要更新，及從篩中將主要資料元件刪除。另一方法是將對於該主要資料元件所駐存的葉狀節點中之主

要資料元件的條目之參照傳達至所有的地點。

【0192】因此，資料提取(Data Distillation™)設備可被用來縮減橫跨寬域網路的不同地點所儲存之資料的足跡，並且使網路之通訊鏈路的使用有效率。

【0193】第12C至12K圖顯示根據在此所敘述的一些實施例之藉由用於各種使用模型的資料提取(Data Distillation™)設備所產生之縮減資料的各種組件。

【0194】第12C圖顯示資料提取(Data Distillation™)設備1203如何攝取一組輸入檔案1201，以及在完成提取處理之後，如何產生一組提取之檔案1205及主要資料篩或主要資料儲存1206。第12C圖本身之主要資料篩或主要資料儲存1206係由兩個組件所構成，亦即，如第12D圖中所示之映射器1207及主要資料元件(或PDE)1208。

【0195】映射器1207本身具有兩個組件於其中，亦即，一組樹狀節點資料結構及一組葉狀節點資料結構，其界定整體樹狀物。該組樹狀節點資料結構可被放置至一或多個檔案之內。同樣地，該組葉狀節點資料結構可被放置至一或多個檔案之內。在某些實施例中，所謂樹狀節點檔案之單一檔案保存為給定資料集(輸入檔案1201)的主要資料元件建立之樹狀物的整組樹狀節點資料結構，以及所謂葉狀節點檔案之另外的單一檔案保存為該資料集的主要資料元件建立之樹狀物的整組葉狀節點資料結構。

【0196】在第12D圖中，主要資料元件1208包含為給定之資料集(輸入檔案1201)建立的該組主要資料元件。該

組主要資料元件之可被放置至一或多個檔案之內。在某些實施例中，所謂PDE檔案之單一檔案保存為給定資料集建立之整組主要資料元件。

【0197】在樹狀節點檔案中之樹狀節點將包含對該樹狀節點檔案內之其它樹狀節點的參照。在樹狀節點檔案中之最深層(或最下方層次)的樹狀節點將包含對該葉狀節點檔案中的葉狀節點資料結構中之條目的參照。在葉狀節點檔案中的葉狀節點資料結構中之條目將包含對PDE檔案中之主要資料元件的參照。

【0198】該樹狀節點檔案、葉狀節點檔案及PDE檔案係描繪在第12E圖中，其顯示由該設備所建立出之所有組件的細節。第12E圖顯示一組輸入檔案1201，包含命名為檔案1、檔案2、檔案3、...、檔案N之N個檔案，其係由資料提取(Data Distillation™)設備所縮減，而產生一組提取之檔案1205和主要資料節之各種組件，亦即，樹狀節點檔案1209、葉狀節點檔案1210及PDE檔案1211。提取之檔案1205包含命名為file1.dist(檔案1.提取)、file2.dist(檔案2.提取)、file3.dist(檔案3.提取)、...、fileN.dist(檔案N.提取)之N個檔案。資料提取(Data Distillation™)設備將輸入的資料分解成它的構成物元件，以及建立起兩種類別的資料元件---主要資料元件及衍生物元件。該等提取檔案包含以無損縮減形式之資料元件的說明，且包含對PDE檔案中之主要資料元件的參照。在輸入檔案1201中之各檔案具有在提取之檔案1205中之對應提取之檔案1205。例如，在輸

入檔案 1201 中之檔案 1 1212 對應於提取之檔案 1205 中之命名為檔案 1.提取 1213 之提取檔案。

【0199】應注意的是，第 12E 圖顯示根據第 1A 圖之基於提取檔案及主要資料節之組織而由資料提取 (Data Distillation) 設備所產生的各種組件，其中重建程式係放置在提取檔案中之元件的無損縮減表示之中。應注意的是，某些實施例(根據第 1B 圖)可將重建程式放置於主要資料節中，且對待它們就像主要資料元件一樣。在提取檔案中之元件的無損縮減表示將包含對主要資料節中的重建程式的參照(而非包含重建程式本身)。在這些實施例中，重建程式將如主要資料元件一樣地被對待，以及被產生於 PDE 檔案 1211 中。在又一實施例中，根據第 1C 圖，重建程式係與主要資料元件分離地儲存在所謂重建程式儲存的結構中。在這種實施例中，在提取檔案中之元件的無損縮減表示將包含對重建程式儲存中的重建程式的參照。在這種實施例中，如第 12F 圖所描繪地，除了為主要資料元件之樹狀組織產生的樹狀節點檔案 1209、葉狀節點檔案 1210 及 PDE 檔案 1211 之外，該設備亦將產生第二組之樹狀及葉狀節點檔案(稱作重建樹狀節點檔案 1219 及重建葉狀節點檔案 1220)，連同包含所有重建程式的檔案(稱作 RP 檔案 1221)。

【0200】在第 12E 圖中所示之資料提取 (Data Distillation™) 設備亦儲存支配樹狀節點檔案 1209、葉狀節點檔案 1210、PDE 檔案 1211 及提取之檔案 1205 之一或多者

中之其操作的組態及控制資訊。選擇性地，包含此資訊的第五組件可被產生。相似地，對於第12F圖中所示之設備，組態及控制資訊可被儲存在第12F圖中所示之各種組件的一或多者中，或可被儲存在針對此目的所產生的另一組件中。

【0201】第12G圖顯示資料提取(Data Distillation™)設備之使用的概觀，其中給定之資料集(輸入資料集1221)係饋送至資料提取(Data Distillation™)設備1203並處理，以產生無損縮減之資料集(無損縮減資料集1224)。輸入資料集1221可由檔案、物件、區塊、資料塊或是來自資料流之提取的集合所構成。應注意的是，第12E圖顯示資料集係由檔案所構成的範例。第12G圖的輸入資料集1221對應於第12E圖的輸入檔案1201，而第12G圖的無損縮減資料集1224包含第12E圖中所示之四個組件，亦即，第12E圖之提取之檔案1205、樹狀節點檔案1209、葉狀節點檔案1210及PDE檔案1211。在第12G圖中，資料提取(Data Distillation™)設備利用橫跨所提交至其之輸入資料集的整個範圍之資料元件中的冗餘。

【0202】資料提取(Data Distillation™)設備可被組構以利用橫跨輸入資料集的子集的冗餘，且傳遞用於所提交至其之資料的各子集之無損縮減。例如，如第12H圖中所示，輸入資料集1221可被劃分成許多較小的資料之集合，各集合係在此發明中稱作“批量”或“資料之批量”或“資料批量”。第12H圖顯示被組構以攝取輸入資料批量1224及產

生無損縮減資料批量 1225 的資料提取 (Data Distillation™) 設備。第 12H 圖顯示由許多資料集合所構成之輸入資料集 1221，該等資料集合係資料批量 1、...、資料批量 i、...、資料批量 n。資料係每次提交一個資料批量至資料提取 (Data Distillation™) 設備，且冗餘係橫跨各資料批量的範圍而被利用，用以產生無損縮減資料批量。例如，來自輸入資料集 1221 的資料批量 i 1226 係饋送至設備，而無損縮減資料批量 i 1228 係傳遞至無損縮減資料集 1227。來自輸入資料集 1221 的各資料批量係饋送至設備，而對應的無損縮減資料批量係傳遞至無損縮減資料集 1227。一旦消耗及縮減所有的資料批量 1、...、資料批量 i、...、資料批量 n，輸入資料集 1221 就可被縮減成無損縮減資料集 1227。

【0203】雖然資料提取 (Data Distillation™) 設備係藉由在利用橫跨資料的全域範圍之冗餘已經有效的設計，但上述技術可被用來進一步加速資料縮減處理，且進一步增進其效率。資料縮減處理的產出量可藉由限制資料批量之大小成為能適合系統的可用記憶體來增加。例如，在大小上係數萬億個位元組或甚至千兆個位元組的輸入資料集可各被拆散成如 256GB 之大小的許多個資料批量，且各資料批量可被快速地縮減。使用具有 256GB 之記憶體的單一處理器核心 (Intel Xeon E5-1650 V3, Haswell 3.5Ghz 處理器)，利用橫跨 256GB 的範圍的這種解決方法已被實施於我們的實驗室中，以產生每秒數百個百萬位元組的攝取速率，且同時對於各種資料集傳遞 2 至 3 倍的縮減層次。應注

意的是，256GB的範圍係大於32KB數百萬倍，所述32KB係窗口的大小，其中Lempel Ziv方法對於當代的處理器在該大小傳遞10MB/秒至200MB/秒之間的攝取效能。因此，藉由適當地限制冗餘的範圍，在資料提取處理之速度中的增進可藉由潛在地犧牲某些縮減來達成。

【0204】第12I圖顯示第12H圖中之設置的變化例，且顯示運行在多個處理器上之多個資料提取處理，用以顯著地提高輸入資料集之資料縮減(以及，資料重建/檢索)的產出量。第12I圖顯示被劃分成x數目之資料批量的輸入資料集1201，且該x個獨立的資料批量係饋送至運行在獨立的處理器核心上之j個獨立的處理之內(各處理係分配足夠的記憶體以容納將被饋送至它的任一資料批量)，用以並聯地執行且產生大約j倍的加速以供資料縮減以及重建/檢索。

【0205】第12H圖顯示配置成攝取輸入資料批量1224並產生無損地減少的資料批量1225的資料提取(Data Distillation™)設備。第12H圖顯示輸入資料集1221，其係由許多資料集合所構成，該等資料集合係資料批量1、...、資料批量i、...、資料批量n。在一些實施例中，可以採用將輸入資料集劃分為多個資料批量的替代劃分方案，其中資料批量邊界是動態地確定的，以充分利用可用記憶體。可用記憶體既可以用來首先保存所有的樹節點，也可以用來保存所有的樹節點和資料批量的所有葉節點，或者最後可以用來保存所有樹節點、葉節點和所有主要資

料元件。這三種不同的選擇使得設備的替代操作點成為可能。例如，將可用記憶體專用於樹節點可以致使在資料批量中容納更大範圍的資料，但是這需要設備必須在需要時從儲存中擷取葉節點以及相關的主要資料元件，從而致使額外的等待時間。可替換地，致力於使可用記憶體適應樹節點和葉節點兩者將加速提取，但會縮減樹的有效大小，從而縮減資料批量中可容納的資料範圍。最後，使用可用記憶體來保存所有的樹節點、葉節點和主要資料元件將致使最快的提取，但可作為單一範圍支援的資料批量的大小將是最小的。在所有這些實施例中，資料批量將在到達記憶體限制時動態關閉，並且來自輸入資料集的後續檔案成為新資料批量的一部分。

【0206】進一步的改進在於可以提高設備的效率以及加速重建程式。在一些實施例中，單一統一的映射器被用於提取，但不是將主要資料元件保存在單一PDE檔案中，而是將主要資料元件保存在N個PDE檔案中。因此，先前的單一PDE檔案被劃分為n個PDE檔案，每個檔案小於特定的臨限值大小，並且當PDE檔案超過該臨限值大小時(由於主要資料元件的安裝而增長時)，每個分區在提取處理期間被建立。對於每個輸入檔案，藉由查詢映射器來繼續提取，以內容關聯地選擇適於衍生的合適主要資料元件，接著衍生從其駐存的合適的PDE檔案中擷取的合適主要資料元件。每個提取檔案被進一步強化以列出含有特定提取檔案參照的主要資料元件的所有PDE檔案(在n個PDE文件

中)。為了重建特定的提取檔案，只有那些列出的PDE檔案需要被載入或打開而為重組而存取。這具有以下優點：對於單一提取檔案或少數提取檔案的重建，只有那些包含該特定提取檔案所需的主要資料元件的PDE檔案需要被存取或保持活動，而其它PDE檔案不需要被保留或者載入到記憶體或儲存器的快速層中。因此，重建可以加快並且更加有效率。

【0207】將PDE檔案劃分為n個PDE檔案可以藉由標準來進一步引導，該標準在資料集中的任何給定檔案的縮減期間使參照模式對主要資料局部化。所述設備可以用計數器來強化，該計數器計數和估計對於當前PDE檔案中的元件的參照密度。如果這個密度很高，則PDE檔案將不會被劃分或拆分，並且在隨後安裝元件時將繼續增長。一旦來自給定提取檔案的參照密度逐漸減少，則PDE檔案就可以被允許在隨後的增長超過特定臨限值時被分裂和劃分。一旦被劃分，新的PDE檔案將被打開，並且來自隨後提取的隨後安裝將被製成這個新的PDE檔案。這種安排將進一步加快重建，如果只有來自資料批量的一部分檔案需要重建。

【0208】第12J圖顯示由資料提取(Data Distillation™)設備所產生之用於使用模型之所縮減資料的各種組件，其中在輸入資料集的縮減之後，映射器無需再被保持。這種使用模型的範例係某些種類的資料備份及資料歸檔的應用。在這種使用模型中，所縮減資料的隨後使用係從所縮

減資料集的輸入資料集之重建和檢索。在這種情況中，所縮減資料的足跡可藉由在完成資料縮減之後不儲存映射器來進一步縮減。第12J圖顯示被饋送至設備的輸入檔案1201，其產生提取之檔案1205和PDE檔案1211---在此情況中，這些組件包含縮減之資料。應注意的是，輸入檔案1201可僅使用提取之檔案1205和PDE檔案1211來完全地再生和復原。回顧一下，在提取檔案中用於各元件之無損縮減表示包含在該處所需的重建程式，以及對PDE檔案中之主要資料元件的參照。與PDE檔案相結合，這是執行重建所需的所有資訊。同樣值得注意的是，指出這種安排對重建和檢索輸入資料集的效能效率的重要益處。在這個實施例中，該設備將輸入資料集分解為提取檔案和包含在單獨的PDE檔案中的主要資料元件。在重建期間，可以首先將PDE檔案從記憶體載入到可用記憶體中，並且隨後可以從儲存器中連續讀取提取檔案以供重建。在重建每個提取檔案期間，重建提取檔案所需的任何主要資料元件將快速從記憶體中檢索，而不會致使讀取主要資料元件的任何額外儲存存取延遲。重建的提取檔案可以在完成時寫到儲存器。這種安排排除了執行隨機儲存存取的需要，否則會對效能產生有害影響。在此解決方案中，載入來自儲存器的PDE檔案的是一組對於順序連續的位元組資料塊的存取，每個提取檔案的讀取也是一組對於順序連續的位元組資料塊的存取，並且最後每個重建的輸入檔案被寫出到儲存器作為一組順序連續的位元組資料塊的存取。這種安排的儲

存效能更緊密地追蹤順序連續讀取和寫入位元組資料塊的效能，而不是致使多個隨機儲存存取的解決方案的效能。

【0209】應注意的是，第12J圖顯示根據第1A圖之基於提取檔案及主要資料節之組織而由資料提取(Data Distillation)設備所產生的各種組件，其中重建程式係放置在提取檔案中之元件的無損縮減表示之中。應注意的是，某些實施例(根據第1B圖)可將重建程式放置在主要資料節中，且對待它們就像主要資料元件一樣。在提取檔案中之元件的無損縮減表示將包含對主要資料節中的重建程式的參照(而非包含重建程式本身)。在這些實施例中，重建程式將如主要資料元件一樣地被對待，以及被產生於PDE檔案1211中。在又一實施例中，根據第1C圖，重建程式係與主要資料元件分離地儲存在所謂重建程式儲存的結構中。在這種實施例中，在提取檔案中之元件的無損縮減表示將包含對重建程式儲存中的重建程式的參照。在這種實施例中，除了產生用於主要資料元件之PDE檔案之外，該設備亦將產生包含所有重建程式的檔案(稱作RP檔案)。此係顯示於第12K圖中，其顯示用於使用模型之所縮減資料的組件，其中映射器無需再被保持。第12K圖顯示包含提取之檔案1205、PDE檔案1211及RP檔案1221的縮減資料組件。

【0210】第12L至P圖根據在此描述的一些實施例顯示提取處理可以在分散式系統上如何部署和執行，以便能夠在非常高的攝取率容納非常大的資料集。

【0211】分散式計算模式意味著藉由在多台電腦上運行的程式的大型資料集的分散式處理。第12L圖顯示在稱作分散式計算集群的組織中聯網在一起的一些電腦。第12L圖顯示在電腦之間的點對點鏈路，但應當理解，可以使用任何通訊拓撲，例如，中心輻射拓撲或網狀拓撲來代替第12L圖中所示的拓撲。在給定的群集中，一個節點被指定為將任務分配給從屬節點且控制並協調他們的整體操作的主要節點。從屬節點執行主要節點所指示的任務。

【0212】資料提取處理可以跨分散式計算集群的多個節點用分布的方式來執行以利用集群中的許多電腦的總計算、記憶體和儲存容量。在這種設置中，在主要節點上的主要提取模組與在從屬節點上運行的從屬提取模組互動以實施分散方式的資料提取。為了促進此提取，該設備的主要資料篩可被劃分成多個獨立的子集或子樹，其可以跨運行在從屬節點上的多個從屬模組來分布。回顧一下，在資料提取設備中，主要資料元件根據他們的名稱而以樹的形式來組織，並且他們的名稱係從他們的內容而衍生。主要資料篩可以根據主要資料篩中的元件名稱的前導位元組被劃分成多個獨立的子集或子篩。可以有多種方式來劃分跨多個子樹的名稱空間。例如，元件的名稱的前導位元組的值可以被劃分成多個子範圍，而每個子範圍被分配給子篩。可以有跟集群中的從屬模組一樣多的子集或分區被建立，所以每個獨立的分區部署在特定的從屬模組上。使用部署的子篩，每個從屬模組被設計來對於它所接收的候選

元件執行資料提取處理。

【0213】第12M圖顯示劃分成將被部署在運行於4個節點的4個從屬模組之標記為PDS_1、PDS_2、PDS_3和PDS_4的4個主要資料篩或子篩的主要資料篩的樣本。該劃分是根據主要資料元件的名稱的前導位元組。在所示的範例中，在PDS_1中的所有元件的名稱的前導位元組將在範圍A至I中，並且篩PDS_1將具有由引導到它的值的範圍所標記的名稱A_I。同樣地，在PDS_2中的所有元件的名稱的前導位元組將在範圍J至O中，並且子篩PDS_2將具有由引導到它的值的範圍所標記的名稱J_O。同樣地，在PDS_3中的所有元件的名稱的前導位元組將在範圍P至S中，並且子篩PDS_3將具有由引導到它的值的範圍所標記的名稱P_S。最後，在PDS_4中的所有元件的名稱的前導位元組將在範圍T至Z中，並且子篩PDS_4將具有由引導到它的值的範圍所標記的名稱T_Z。

【0214】在此設置中，在主要節點上運行的主要模組接收輸入檔案，並且進行輸入檔案的輕量解析和分解以將輸入檔案分切成候選元件序列，並且隨後引導每個候選元件到合適的從屬模組以供進一步處理。輕量解析可以包含對一種模式解析每個候選元件，或者可能包含對候選元件的指紋圖譜的應用，以確定構成候選元件的名稱的前導位元組的維度。在主要節點的解析限於只識別足以確定哪個從屬模組應該接受候選元件的位元組數。根據在候選元件的名稱的前導位元組的值，該候選被轉發到在保持對應於

此特定值的子篩的從屬節點的從屬模組。

【0215】當資料累積到篩中時，分區可以間歇地重新訪視和重新平衡。分區和重新平衡功能都可以藉由主要模組來執行。

【0216】在接收候選元件時，每個從屬模組執行資料提取處理，從候選元件的完整解析和檢查開始，以建立其名稱。使用這個名稱，從屬模組執行子篩的內容關聯查找，並執行提取處理以將候選元件轉換成關於那個子篩的無損縮減表示中的元件。在提取檔案中的元件的無損縮減表示以稱為從屬數(SlaveNumber)的欄位來強化以識別從屬模組以及關於哪個元件已經被縮減的對應子篩。元件的無損縮減表示被發送回主要模組。如果在子篩中找不到候選元件，或不能從在子篩中的主要資料元件得到，新的主要資料元件被識別為以分配到子篩。

【0217】主要模組繼續以引導來自輸入檔案中的所有候選元件至適當的從屬模組，並累計收到的元件描述(以無損縮減表示)，直到它已經收到輸入檔案的所有元件。在這一點上，全域承諾通訊可以發給所有從屬模組以利用他們個別的提取處理的結果來更新他們各自的子篩。用於輸入的提取檔案被儲存在主要模組。

【0218】在一些實施例中，不是在任何從屬可以利用新的主要資料元件或元資料來更新其子篩之前等待整個提取檔案準備好，而是當候選元件在從屬模組接受處理時，可以完成對於子篩的更新。

【0219】 在一些實施例中，每個子篩包含按照針對第1B和1C圖描述的主要資料元件以及重建程式。在這種實施例中，重建程式被儲存在子篩中，並且無損縮減表示包含對於子篩中的主要資料元件以及重建程式(如有必要)兩者的參照。這進一步縮減了元件的大小，從而縮減了需要被儲存在主要模組的提取檔案的大小。在一些實施例中，每個子篩中的主要重建程式篩包含那些用於建立駐存在該子篩中的主要資料元素的衍生物的重建程式。在這種情況下，主要重建程式可在從屬節點區域地獲得，並且能夠快速衍生和重建，而不會有任何延遲，否則將會從遠端節點擷取主要重建程式。在其它實施例中，主要重建程式篩橫跨所有的節點全域地分布以利用分散式系統的總容量。無損縮減表示係藉由第二欄位來增強，該第二欄位識別包含主要重建程式的從屬節點或子篩。在這種實施例中，解決方案致使額外的延遲以從遠端節點擷取主要重建程式，以便藉由衍生或者重建元建來產生最終主要重建程式。整體方法利用所有的從屬節點的組合儲存容量來根據每個組塊或每個檔案中的候選元件的內容來橫跨所有節點來分發檔案。

【0220】 資料檢索同樣由主要模組協調。主要模組接收提取檔案，並檢查提取檔案中的每個元件的無損縮減規範。其提取了指示哪個從屬模組將重建元件的欄位“從屬數(SlaveNumber)”。該元件接著被發送給適當的從屬模組以供重建。重建的元件接著被發送回主要模組。主要模組

匯集來自所有從屬的重建元件，並且將所重建的檔案轉發到正在要求該檔案的使用者。

【0221】第12N圖顯示資料提取設備在分散式系統中可如何被部署和執行。輸入檔案1251被饋送到解析和識別檔案中的每個候選元件的名稱的前導位元組的主要模組。主要模組將候選元件轉向4個從屬模組中之一者。保有PDS_1或具有包含承載範圍A至I的值的名稱的前導位元組的主要資料元件的名稱A_I的子篩的從屬節點1處的從屬模組1接收了具有名稱BCD...的候選元件1252，其被確定為已經存在於具有名稱A_I的子篩中的元件的副本。從屬模組1返回無損縮減表示1253，其包含該元件是主要的且駐存在位址refPDE1處的SLAVE1中的指示符。如第12N圖所示，該主要模組將所有候選元件發送到相關的從屬模組，並且組裝和收集，最後儲存該提取檔案。

【0222】第12O圖顯示在第12N圖中所示的方案的變化。在這種變化中，在提取檔案中的每個元件的無損縮減表示中，識別相對於已縮減的該元件的特定Child_Sieve的欄位包含Child_Sieve的名稱，而不是Child_Sieve所在的模組或節點的數目。因此，該欄位SlaveNumber由欄位Child_Sieve_Name取代。這具有藉由其虛擬位址對於相關Child_Sieve參照的好處，而不是Child_Sieve所在的模組或實體節點的數目。因此，如可從第12O圖中看出，保有PDS_1或具有包含承載範圍A至I的值的名稱的前導位元組的主要資料元件的名稱A_I的子篩的從屬節點1處的從屬模

組 1 接收了具有名稱 BCD... 的候選元件 1252，其被確定為已經存在於具有名稱 A_I 的子篩中的元件的副本。從屬模組 1 返回包含該元件是主要的且駐存在位址 refPDE1 處的具有名稱 A_I 的 Child_Sieve 中的指示符的無損縮減表示 1254。

【0223】注意，藉由採用在第 12L 至 12O 圖中描述的佈置，資料提取處理的整體產出速率可以增加。在主要模組處的產出量現在將藉由來自主要模組的候選元件之輕量解析和調度來限定。眾多候選元件的提取將並行執行，只要他們的內容轉向到不同的從屬模組。

【0224】為了進一步提高總產出量，用來識別哪些 Child_Sieve 應接收候選元件之輸入流的輕量解析和分解的任務可以並行。此任務可由主要模組劃分成藉由運行在多個從屬節點上的從屬模組來並行執行的多個並行任務。這可以藉由在資料流中展望和將資料流切片成多個部分重疊的片段來實施。這些片段由主要模組發送到並行執行輕量解析和分解並發送回分解的結果到主要模組的從屬模組中的每一個。主要模組解決橫跨每個片段的邊界的分解，接著將候選元件路由到合適的從屬模組。

【0225】第 12L 至 12O 圖描述了資料提取設備利用在主要節點上運行的主要提取模組和在從屬節點上運行的多個從屬提取模組以分散式的方式操作的佈置。主要模組負責執行橫跨各種子篩的主要資料元件的分區。在所示的佈置中，所有將被攝取的輸入檔案是由主要模組攝取，並且

無損減縮提取檔案被保留在主要模組，而所有主要資料元件(和任何主要重建程式)駐存在各種從屬模組處的子篩中。對於檔案的資料檢索請求也由該主要模組處理，並且對應的提取檔案的重建由該主要模組協調。第12P圖顯示輸入檔案可藉由任何從屬提取模組攝取(和對應的提取檔案保持在那些模組)的變型例，並且資料檢索請求可以由任何從屬提取模組來處理。主要模組以相同方式橫跨子篩繼續執行主要資料元件的分割，使得橫跨子篩的主要資料元件的分布將相同於第12L至12O圖中所示的佈置。然而，在第12P圖中所示的新佈置中，每個從屬模組意識到分區，因為每個從屬模組可以攝取和檢索資料。此外，所有的模組都意識到由那些模組在資料的攝入時在每個模組處建立和儲存的提取檔案的存在和位置。這使得任何從屬模組滿足對於儲存在整個系統中的任何檔案的資料檢索請求。

【0226】如第12P圖所示，每個從屬模組可攝取並檢索來自分散式儲存系統的資料。例如，從屬提取模組1 1270攝取了輸入檔案I 1271，並執行輕量解析以將輸入檔案I分解並將候選元件路由到包含對應於來自輸入檔案I的每個候選元件的名稱的子篩的模組。例如，來自輸入檔案I的候選元件1275被送到從屬提取模組2 1279。同樣地，從屬提取模組2 1279攝取了輸入檔案II，並執行輕量解析以將輸入檔案II分解並將候選元件路由到包含對應於來自輸入檔案II的每個候選元件的名稱的子篩的模組。例如，來

自輸入檔案II的候選元件1277被送到從屬提取模組1 1270。每個從屬提取模組處理他們所收到的候選元件、完成關於他們的子篩的提取處理，並將候選元件的無損縮減表示返回到攝取資料的發起模組。例如，響應於從來自從屬提取模組1 1270的輸入檔案I接收候選元件1275，從屬提取模組2 1279將無損縮減元件1276返回到從屬提取模組1 1270。同樣地，響應於從來自從屬提取模組2 1279的輸入檔案II接收候選元件1277，從屬提取模組1 1270將無損縮減元件1278返回到從屬提取模組2 1279。

【0227】在這種佈置中，在任何從屬模組處可以滿足檢索的資料。接收檢索請求的模組需要首先確定用於請求檔案的提取檔案駐存之處，以及從對應的從屬模組擷取提取檔案。接著，啟動從屬模組需要協調在該提取檔案中的各種元件的分布重建以得到原始的檔案並將其傳遞給請求的應用程式。

【0228】以這種方式，資料提取處理可以用分散式的方式橫跨分散式系統的多個節點來執行以更有效地利用在集群中的許多電腦的總計算、記憶體和儲存容量。系統中的所有節點可用於攝取和檢索資料。這應致使非常高速率的資料攝取和檢索，同時充分利用在系統中的節點的總合儲存容量。這也使得在系統中的任何節點上運行的應用程式來對於儲存在系統中任何地方的任何資料的區域節點進行查詢，並具有效率地和無縫地滿足的查詢。

【0229】在第12M至12P圖中所述的佈置中，橫跨在

該系統의 各個節點中駐存의 子篩의 資料의 劃分是根據在全域可見의 名稱空間中の 元件의 名稱，其中元件係藉由分解輸入檔案來提取。在另一佈置中，共享某些元資料의 資料批量或整組의 檔案可以被分配並儲存在特定の節點。因此，總體資料의 主要分區係根據資料批量，並且由主要模組來執行和管理。所有從屬模組保持意識到資料批量對於模組의 分配。資料批量將完全駐存在給定の從屬節點上。在運行於該從屬節點上の提取從屬模組上の子篩將包含屬於這個資料批量의 所有主要資料元件。換句話說，給定資料批量의 所有主要資料元件의 整個樹將完全駐存在單一從屬提取模組內的單一子篩上。給定資料批量의 所有提取檔案也將駐存在同一從屬提取模組上。使用這種佈置，輸入檔案仍然可以由任何從屬提取模組來攝取，而資料檢索請求仍然可以藉由任何從屬提取模組來處理。然而，給定資料批量의 整個資料提取處理完全在包含該資料批量의 模組上執行。針對資料攝取和資料檢索의 請求係從發起模組路由到被指定持有特定資料批量의 特定從屬模組。當分解和提取資料批量時，這種解決方案具有降低在分散式環境中の通訊額外成本의 好處。冗餘不再橫跨整個全域資料足跡被利用，但非常有效地在資料批量內局部利用。這種解決方案仍然使用分散式系統의 組合儲存容量，並提供從系統의 任何節點中查詢、攝取和檢索任何資料의 無縫能力。

【0230】因此，採用上述의 眾多技術，有效地利用在分散式系統中の資源以用非常高的速度對於非常大的資料

集執行資料提取。

【0231】可進一步強化資料提取(Data Distillation™)方法和設備，以利於資料的有效移動和遷移。在一些實施例中，無損縮減的資料集可以用多個容器或包裹的形式來傳遞以利於資料移動。在一些實施例中，一或多個縮減的資料批量可以適合於單一容器或包裹，並且可選地，單一縮減的資料批量可以被轉換為多個包裹。在一些實施例中，單一縮減的資料批量作為單一自描述包裹來遞送。第12Q圖說明了這種包裹的樣本結構。第12Q圖中的包裹1280可被視為單一檔案或連續的一組位元組，其包含順序地彼此串聯的以下組件：(1)標頭1281，其是首先包含指定包裹長度的包裹長度1282的包裹標頭，接著包含偏移識別符，其用於識別在包裹中的提取檔案、PDE檔案和各種清單的偏移量；(2)提取之檔案1283，其為用於一個接一個地連接在一起的資料批量的提取檔案，其中首先指定每個提取檔案的長度，接著是所有包含提取檔案的位元組；(3)PDE檔案1284，其為PDE檔案，從PDE檔案的長度識別符開始，接著是包含所有主要資料元件的PDE檔案的主體；(4)來源清單1285，其為描述輸入資料集的結構，並識別包裹中每個檔案的唯一目錄結構、路徑名稱和檔案名稱的來源清單。來源清單還包含輸入資料批量中的每個節點(已縮減並轉換為包裹)的列表以及與每個節點關聯的元資料；(5)目的地清單和映射器1286，其為目的地清單和映射器。目的地映射器將每個輸入節點和檔案的預期映射

提供到目標目的地目錄和檔案結構或雲端中的目標桶/容器和物件二進制大型物件(blob)結構。此清單有助於在資料移動之後將包裹中各個組件移動、重建和重新定位到最終目的地。請注意，可以獨立更改此目的地映射器部分，以重新定向包裹中的資料要傳輸到並重建的目的地。

【0232】藉由這種方式，資料批量的無損縮減表示是作為以自描述的格式的包裹遞送，其適合於資料的移動和遷移。

【0233】資料縮減係使用在此所敘述之實施例而對於各種真實世界的資料集執行，用以決定該等實施例的效用。真實世界所研究的資料集包含企業電子郵件的安隆語料庫(Enron Corpus)、各種美國政府記錄及文獻、進入至MongoDB NOSQL資料庫之美國運輸部門記錄及提供給公眾的企業PowerPoint演講稿。使用在此所敘述之實施例，且將輸入的資料分解成平均4KB之可變大小的元件(具有由指紋圖譜所決定的邊界)，3.23倍之平均資料縮減係在該等資料集的範圍取得。3.23倍之縮減意指的是，所縮減資料之大小係等於原始資料的大小除以3.23倍，而導致具有31%的壓縮比之縮減的足跡。傳統的重複資料刪除技術被發現使用等效的參數而在該等資料集上傳遞1.487倍之資料縮減。使用在此所敘述之實施例，且分解輸入的資料成平均4KB之固定大小的元件，1.86倍之平均資料縮減係在該等資料集的範圍取得。傳統的重複資料刪除技術被發現使用等效的參數而在該等資料集上傳遞1.08倍之資料縮

減。因此，資料提取(Data Distillation™)解決方法被發現比傳統重複資料刪除解決方法傳遞顯著更佳的資料縮減。

【0234】測試運行亦證實的是，主要資料元件之小的子集之位元組用以排列篩中的多數的元件，而藉以致能用於其操作所需之最小增量儲存的解決方法。

【0235】結果證實資料提取(Data Distillation™)設備有效率地致能以比元件本身更細的粒度來全域地橫跨整個資料集利用資料元件中的冗餘。由此方法所傳遞的無損資料縮減係以經濟的資料存取和IO，採用其本身所需之最小增量儲存的資料結構，以及使用現代多核心微處理器可得之總計算處理能力的一部分來達成。在之前的段落中所敘述之實施例賦予系統及技術之特徵，該等系統及技術之特徵在於，對於大且極大的資料集執行無損資料縮減，且同時提供高速率的資料攝取和資料檢索，而不會遭遇到習知技術的缺點及限制。

對於已經藉由從駐存在主要資料篩中的主要資料元件衍生資料來無損縮減的資料執行內容關聯的搜索和檢索

【0236】在前面的文字中所描述和第1A至12P圖中所顯示的資料提取設備可以利用某些特徵來強化，以便有效地對於來自以無損縮減形式儲存的資料的資訊執行多維度搜索與內容關聯檢索。這種多維度搜索和資料檢索是分析或資料倉儲應用的關鍵構建模組。現在將描述這些強化。

【0237】第13圖顯示類似於第3H圖中所示結構的葉節

點資料結構。然而，在第13圖中，在用於每個主要資料元件的葉節點資料結構中的條目被強化，以包含對於含有對於特定主要資料元件的參照的提取資料中的所有元件的參照(這也將被稱為反向參照或反向鏈路)。回顧一下，該資料提取方案將來自輸入檔案的資料分解成元件的序列，該些元件的序列配置在使用如第1H圖所描述的規範之縮減形式中的提取檔案中。在提取檔案中有兩種元件：主要資料元件和衍生物元件。針對在提取檔案中的這些元件中的每一個的規範都將包含對於駐存在主要資料篩中的主要資料元件的參照。對於(來自提取檔案中的元件對於主要資料篩中的主要資料元件的)這些參照中的每一個會有對應的反向鏈路或反向參照(從用於葉節點資料結構中的主要資料元件的條目到提取檔案中的元件)被安裝在葉節點資料結構中。反向參照決定了標記元件的無損縮減表示的起點的提取檔案內的偏移量。在一些實施例中，反向參照包含提取檔案的名稱和定位元件的起點的該檔案內的偏移量。如第13圖所示，隨著對於在提取檔案中的每個元件的反向參照，葉節點的資料結構也保有識別在提取檔案中正在參照的元件是否為主要資料元件(**prime**)，或它是否為衍生物元件(**deriv**)的指示符。在提取處理期間，如果且當元件被放入該提取檔案中時，則反向鏈路被安裝到該葉節點資料結構。

【0238】反向參照或反向鏈路被設計為一種可以接觸共用主要資料篩的所有提取檔案中的所有元件的通用處

理。

【0239】因為資料元件大小預期被選擇，反向參照的添加預計不會顯著影響所完成的資料縮減，使得每個參照是資料元件的大小的一部分。例如，考慮一系統，其中衍生物元件被限制為每個衍生不超過1個主要資料元件(因此，多元件衍生是不允許的)。橫跨所有葉節點資料結構的反向參照的總數將等於橫跨所有提取檔案的元件的總數。假設32GB大小的樣本輸入資料集被縮減到8GB的無損縮減資料(採用1KB的平均元件大小)並產生4X的縮減比。在輸入資料中有32M個元件。如果每一個反向參照的大小為8B，反向參照所佔據的總空間為256MB或0.25GB。對於8GB足跡的縮減資料這是小的增加。新的足跡將是8.25GB，並且完成的有效縮減將是3.88X，其代表3%的縮減損失。對於縮減資料的強大內容關聯資料檢索的好處，這是很小的代價。

【0240】如先前在本文中所描述的，提取設備可以採用多種方法來確定候選元件的內容之內的骨架資料結構的各種組件的位置。元件的骨架資料結構的各種組件可以被視為維度，使得跟隨著各元件的內容的其餘部分的這些維度的串聯係用於建立每個元件的名稱。該名稱用於排序和組織樹中的主要資料元件。

【0241】在輸入資料的結構是已知的使用模型中，一種模式定義了各種欄位或維度。這種模式是由使用此內容關聯資料檢索設備的分析應用程式來提供，並藉由到應用

程式的介面來提供給該設備。根據在模式中的宣告，提取設備的解析器能夠解析候選元件的內容以檢測和定位各種維度和建立候選元件的名稱。如先前所述，在對應於該維度的欄位中具有相同內容的元件將沿著樹的同一分支被群組在一起。對於安裝到篩中的每個主要資料元件，維度上的資訊可以被儲存為用於葉節點資料結構中的主要資料元件的條目中的元資料。此資訊可以包含在每個在所宣告維度的內容的位置、大小和值，並可以被儲存在第13圖中提到的欄位為“主要資料元件的其它元資料”。

【0242】第14A圖顯示根據在此描述的一些實施例的提供輸入資料集的結構的描述和輸入資料集的結構與維度之間的對應關係的描述的範例模式。結構描述1402是描述輸入資料的完整結構的更完整的模式的一段或一部分。結構描述1402包含一連串的關鍵字（如“PROD_ID”、“MFG”、“MONTH”、“CUS_LOC”、“CATEGORY”和“PRICE”），其次是對應於關鍵字的值的類型。冒號“：”作為分隔符來將關鍵字與值的類型分開，並且分號“；”作為分隔符將不同對的關鍵字和值的對應類型分開。需要注意的是，（結構1402為一部分的）完整的模式可以指定額外的欄位來識別每一個輸入的開始和結束，並且也可能是維度以外的其它欄位。維度映射描述1404描述了用於組織主要資料元件的維度如何映射成結構化的輸入資料集中的關鍵字值。例如，在維度映射描述1404中的第一行指定對應於在輸入資料集中的關鍵字“MFG”的值得前四個位元組（因

為第一行以文字“前綴=4”結尾)被用來產生維度1。在維度映射描述1404中的剩下的行描述了如何根據結構化的輸入資料來建立其它三個維度。在這種關鍵字到維度的映射中，作為他們出現在輸入中的關鍵字的順序不一定匹配維度的順序。使用提供的模式描述，解析器可以辨識在輸入資料中的這些維度以建立候選元件的名稱。對於第14A圖中的範例，並使用維度映射描述1404，候選元件的名稱將被建立如下：(1)名稱的前4個位元組將是來自對應於如維度1所宣告的關鍵字“MFG”的值的的前4個位元組，(2)名稱的接下來的4個位元組將是來自對應於如維度2所宣告的關鍵字“CATEGORY(類別)”的值的的前4個位元組，(3)名稱的接下來的3個位元組將是來自對應於如維度3所宣告的關鍵字“CUS_LOC”的值的的前3個位元組，(4)名稱的接下來的3個位元組將是來自對應於如維度4所宣告的關鍵字“MONTH(月)”的值的的前3個位元組，(5)名稱的下一組位元組將由來自維度的其餘位元組的串聯所組成，(6)而最後，在維度的所有位元組都用盡之後，名稱的其餘位元組將從候選元件的其餘位元組的串聯來建立。

【0243】藉由驅動這種設備的應用程式所提供的模式可以指定一些主要維度，以及一些次要維度。用於所有這些主要和次要維度的資訊可以被保留在葉節點資料結構中的元資料中。主要維度係用來形成沿著用以將篩中的元件進行排序和組織的主軸線。如果主要維度被用盡，並且具有龐大成員的子樹仍然存在，那麼次要維度也可以用來更

深入樹中以進一步將元件細分成更小的群體。次要維度上的資訊可被保留作為元資料，且也用作為次要標準來區分葉節點內的元件。在提供內容關聯的多維度搜索和檢索的一些實施例中，所有輸入的資料必須包含用於由模式宣告的每個維度的關鍵字和有效值之要求可被配置。這允許系統以一種方式確保只有有效的資料進入篩中的所需子樹。不含有指定為維度的所有欄位或包含對應於用於維度的欄位的值中的無效值的候選元件將被發送到如早先在第3E圖所示的不同子樹。

【0244】資料提取設備係以一種額外的方式來約束，以根據維度中的內容來全面支援內容關聯搜索和檢索。當從主要資料元件建立了衍生物元件，該衍生物器被約束以確保主要資料元件和衍生物在每個對應維度的值欄位中都具有完全相同的內容。因此，當正在建立衍生物時，該重建程式不允許被擾亂或修改對應於主要資料元件的任何維度的值欄位中的內容，以便構建衍生物元件。在篩的查找期間，給定候選元件，如果候選元件相較於目標主要資料元件的對應維度，在任何維度中具有不同的內容，則新的主要資料元件需要被安裝，而不是接受衍生物。例如，如果主要維度的子集足以將元件排序到在樹中不同的群組，以使候選元件到達葉子節點以搜索在主要維度的此子集中具有相同內容，但在其餘主要維度或次要維度中具有不同內容的主要資料元件，接著，新的主要資料元件需要安裝，而不是建立衍生物。此特徵確保了所有資料都可以藉由使

用維度來簡單查詢主要資料節而進行搜索。

【0245】衍生器可以採用各種實施技術來執行約束條件，即候選元件和主要資料元件必須在每個對應維度的值欄位中具有完全相同的內容。衍生器可以從主要資料元件的骨架資料結構中提取位置、長度和包含對應於維度的欄位的內容的資訊。類似地，此資訊係從解析器/分解器接收或者針對候選元件而計算。接下來，可以比較候選元件和主要資料元件的維度的對應欄位的相等性。一旦確認是相等的，衍生器可以繼續衍生的其餘部分。如果不相等，則候選元件作為新的主要資料元件安裝在篩中。

【0246】上述限制，預計不會顯著妨礙大多數使用模型的資料縮減程度。例如，如果輸入資料由每次1000個位元組大小的資料倉儲處理的一組元件組成，並且如果一組6個主要維度和14個次要維度係由模式指定，其每個中的每個維度具有8個位元組的資料，在該維度由內容所佔用的總位元組為160個位元組。當建立衍生物時，不允許對於這些160個位元組擾動。這仍然留下可用於擾動的候選元件資料的剩餘840個位元組來建立衍生物，因此留有充分的機會來開發冗餘，同時致使將來自資料倉儲的資料使用該維度以內容關聯的方式進行搜索與檢索。

【0247】為了執行查詢包含在維度中的欄位的特定值的資料的搜索，該設備可以遍歷樹，並到達符合規定的維度的樹中的節點，並且可以將該節點下的所有葉節點的資料結構返回作為查找的結果。對於存在於葉節點的主要資

料元件的參照可用於在需要時擷取所需的主要資料元件。如果需要的話，反向鏈路致使來自提取檔案之(無損縮減形式的)輸入元件的檢索。元件隨後可以被重建以產生原始輸入資料。因此，強化的設備允許對於在主要資料篩中的資料(其為總資料的較小子集)完成所有搜索，同時也能夠根據需要達到並檢索所有衍生物元件。

【0248】如強化的裝置可用於根據在由查詢所指定的維度中的內容來執行用於資料的相關子集的強大搜索和檢索的搜索和查找查詢。內容關聯資料檢索查詢將具有“擷取(維度1，維度1的值；維度2，維度2的值；...)”的形式。查詢將指定參與搜索的維度以及將用於內容關聯搜索和查找的指定維度中的每一個的值。查詢可以指定所有維度或可以僅指定維度的子集。所述查詢可以根據作為搜索和檢索的標準的多個維度來指定複合條件。在具有針對指定維度的指定值的篩中的所有資料將被檢索。

【0249】多種擷取查詢可以被支援並提供給正在使用此內容關聯資料檢索設備的分析應用程式。這種查詢將藉由介面從應用程式被提供給該設備。該介面提供了來自應用程式對於該設備的查詢，並且從該設備將查詢的結果返回到應用程式。首先，查詢FetchRefs可以被用來將參照或處理擷取到第13圖中的葉節點資料結構(連同子ID或條目的索引)以用於匹配於該查詢的每個主要資料元件。第二種形式的查詢FetchMetaData可以用來從第13圖中的葉節點資料結構中的條目擷取元資料(包含骨骼資料結構、維

度上的資訊，以及對於主要資料元件的參照)以用於匹配於該查詢的每個主要資料元件。第三種形式的查詢 **FetchPDEs** 將擷取符合搜索條件的所有主要資料元件。另一種形式的查詢 **FetchDistilledElements** 將擷取符合搜索條件的提取檔案中的所有元件。又一種形式的查詢 **FetchElements** 將擷取符合搜索條件的輸入檔案中的所有元件。注意，對於 **FetchElements** 查詢，該設備將首先擷取提取元件，接著將有關提取元件重建成來自輸入資料的元件，並且將這些返回作為查詢的結果。

【0250】除了這種多維度內容關聯擷取基元之外，該介面也可以提供給應用程式用以直接存取主要資料元件(使用對於主要資料元件的參照)和提取檔案中的元件(使用對於元件的反向參照)的能力。此外，該介面可以提供給應用程式將提取檔案中的提取元件(給定對於提取元件的參照)重建，並在其存在於輸入資料中時遞送該元件的能力。

【0251】這些查詢的恰當組合可以藉由分析應用程式來使用，以執行搜索、確定有關的關聯和交錯，以及蒐集重要的見解。

【0252】第14B圖說明以下顯示具有在結構描述1402中所描述的結構的輸入資料集的範例。在此範例中，包含在檔案1405中的輸入資料包含電子商務交易。使用第14A圖中的模式和維度宣告，藉由資料提取設備中的解析器來將輸入資料轉換成一連串的候選元件1406。注意，每個候

選元件的名稱的前導位元組如何由來自維度的內容所組成。例如，用於候選元件1的名稱1407的前導位元組是PRINRACQNYCFEB。這些名稱係用於將候選元件組織為樹形式。在資料縮減完成之後，提取資料被放置在提取之檔案1408中。

【0253】第14C圖說明以下顯示維度映射描述1404如何可用於根據結構描述1402來解析第14A圖中顯示的輸入資料集、根據維度映射描述1404來確定維度，並根據確定的維度來將主要資料元件組織為樹。在第14C圖中，主要資料元件組織為使用來自4個維度的總共14個字元的主樹。顯示在主樹中的是各種主要資料元件的葉節點資料結構的一部分。需要注意的是，為了方便查看的目的，並未顯示第13圖的完整葉節點資料結構。然而，第14C圖顯示在葉節點資料結構中的每個條目的路徑資訊或名稱、子ID、從主要資料元件到提取檔案中的元件隨著在提取檔案中的元件是否為“prime”(用P表示)或“deriv”(用D表示)的指示符的所有反向參照或反向鏈路，並且還有對於主要資料元件的參照。第14C圖顯示映射到主樹中的5個主要資料元件的提取檔案中的7個元件。在第14C圖中，用於具有名稱PRINRACQNYCFEB的主要資料元件的反向鏈路A指回到提取檔案中的元件1。同時，具有名稱NIKESHOELAHJUN的主要資料元件有分別到元件2、元件3和元件58的3個反向鏈路B、C和E。需要注意的是，元件3和元件58為元件2的衍生物。

【0254】第14D圖顯示從維度建立的輔助索引或輔助樹以提高搜索的效率。在此範例中，輔助映射樹係根據維度2(其為類別)來建立。藉由直接遍歷此輔助樹，在輸入資料中的給定類別的所有元件可被找到而沒有可能已經發生的主樹的更昂貴遍歷。例如，向下遍歷由“鞋(SHOE)”表示的分支直接致使用於為ADIDSHOESJCSEP和NIKESHOELAHJUN的鞋的兩個主要資料元件。

【0255】可替代地，這種輔助樹可以根據次要維度，並且用於使用該維度來幫助快速收斂搜索。

【0256】現在將提供對於在第14D圖所示的設備執行查詢的範例。查詢FetchPDEs(維度1，NIKE；)將返回名稱為NIKESHOELAHJUN和NIKEJERSLAHOCT的兩個主要資料元件。查詢FetchDistilledElements(維度1，NIKE；)將返回將為無損縮減形式的提取元件的元件2、元件3、元件58和元件59。查詢FetchElements(維度1，NIKE；維度2，SHOE)將返回來自輸入資料檔案1405的交易2、交易3和交易58。查詢FetchMetadata(維度2，SHOES)將返回儲存在用於名為ADIDSHOESJCSEP和NIKESHOELAHJUN的兩個主要資料元件中的每一個的葉節點資料結構條目中的元資料。

【0257】因此，迄今為止描述的設備可用於支援根據在被稱為維度的欄位中指定的內容的搜索。此外，該設備可用於支援根據不包含在維度的列表中的關鍵字列表的搜索。這種關鍵字可以藉由諸如驅動設備的搜索引擎的應

用程式來提供給該設備。關鍵字可以藉由模式宣告被指定到該設備或藉由包含所有關鍵字的關鍵字列表來傳遞，其中每個關鍵字係藉由宣告分隔符來分離(諸如空格或逗號，或換行符號)。可替代地，模式以及關鍵字列表兩者可被用於共同指定所有關鍵字。非常大量的關鍵字可以被指定-設備對於關鍵字的數量沒有任何的限制。這些搜索關鍵字將被稱為關鍵字。該設備可保持用於使用這些關鍵字來搜索的倒置索引。該倒置索引針對每個關鍵字包含含有此關鍵字的提取檔案中的元件的反向參照的列表。

【0258】 根據在模式或關鍵字列表中的關鍵字宣告，提取設備的解析器可以解析候選元件的內容以檢測和定位在輸入的候選元件中的各種關鍵字(如何以及在何處找到)。接著，候選元件藉由資料提取設備被轉換成主要資料元件或衍生物元件並置於如在提取檔案中的元件。在此元件中找到的關鍵字的倒置索引可以隨著對於提取檔案中的此元件的反向參照而被更新。針對在元件中找到的每個關鍵字，倒置索引被更新以包含對於提取檔案中的此元件的反向參照。回顧一下，提取檔案中的元件是以無損縮減表示。

【0259】 當使用關鍵字的資料的搜索查詢時，倒置索引係參照以發現並提取對於包含此關鍵字的提取檔案中的元件的反向參照。使用對於這種元件的反向參照，該元件的無損縮減表示可以被檢索，並且該元件可以被重建。重建的元件可以接著被提供作為搜索查詢的結果。

【0260】倒置索引可以被強化以含有定位重建元件中的關鍵字偏移的資訊。請注意，當對於提取檔案中的元件的反向參照被放置到該倒置索引時，在候選元件中檢測到的每個關鍵字的偏移或位置可由解析器來確定，因此該資訊也可被記錄在倒置索引中。在搜索查詢時，在倒置索引被參照以檢索對於包含相關關鍵字的提取檔案中的元件的反向參照之後，且在該元件被重建之後，在重建元件(如同原始輸入候選元件)中所記錄的關鍵字偏移或位置可以用於查明關鍵字存在於輸入資料或輸入檔案中的何處。

【0261】第15圖顯示用以促進根據關鍵字的搜索的反向索引。對於每個關鍵字，倒置索引包含數值對-第一個值是對於包含關鍵字的提取檔案中的無損縮減元件的反向參照，第二個值是重建元件中的關鍵字的偏移。

【0262】維度和關鍵字對於資料提取設備中的主要資料篩具有不同的含義。請注意，該維度被用作沿著組織篩中的主要資料元件的主軸。維度形成資料中的各元件的骨架資料結構。維度係根據所輸入資料的結構的知識來宣告。該衍生器受到約束，使得所建立的任何衍生物元件必須與對應維度中的每一個的欄位的值中的主要資料元件具有完全相同的內容。

【0263】對於關鍵字，這些屬性不需要保留。既沒有關鍵字甚至存在於資料中的先驗的要求，也沒有主要資料篩必須根據關鍵字來組織，衍生器也不是受限於關於包含關鍵字的內容的衍生物。如果必要，該衍生器可以自由地

藉由修改關鍵字的位置，從主要資料元件來建立衍生物。當掃描輸入資料和反向索引更新時，關鍵字的位置被簡單地記錄在何處發現。在根據關鍵字的內容關聯搜索時，反向索引被查詢並且獲得關鍵字的所有位置。

【0264】在其它實施例中，關鍵字不需要存在於資料中(資料中沒有關鍵字不會使資料無效)，但是主要資料篩需要包含含有關鍵字的所有元件，並且衍生器係受限於關於涉及包含關鍵字的內容的衍生物，除了縮減重複之外，不允許衍生物。這些實施例的目的是包含任何關鍵字的所有不同元件必須存在於主要資料篩中。這是一種範例，其中管理選擇主要資料的規則係由關鍵字決定。在這些實施例中，可以建立修改的反置索引，其對於每個關鍵字包含對於含有關鍵字的每個主要資料元件的反向參照。在這些實施例中，實施了強大的關鍵字式搜索能力，其中僅搜索主要資料篩係與搜索整個資料一樣有效。

【0265】可存在其它實施例，其中衍生器被約束，使得重建程式不允許干擾或修改在主要資料元件中找到的任何關鍵字的內容，以將候選元件制定為主要資料元件的衍生物元件。關鍵字需要從主要資料元件不變地傳播到衍生物。如果衍生器需要修改主要資料元件中找到的任何關鍵字的位元組，以成功地將候選元件制定為此主要資料元件的衍生物，則衍生物可能不被接受，並且候選物必須被安裝為在篩中的新的主要資料元件。

【0266】衍生器可用各種關於涉及關鍵字的衍生物的

方式來限制，使得主管選擇主要資料的規則是由關鍵字來調整。

【0267】使用關鍵字的搜索資料的設備可以接受對於關鍵字列表的更新。關鍵字可以被添加而不將用無損縮減形式儲存的資料作任何修改。當新的關鍵字被添加時，新的輸入資料可針對更新的關鍵字列表進行解析，並且隨著輸入資料更新的倒置索引隨後以無損縮減形式被儲存。如果現有的資料(也就是已經以無損縮減形式被儲存)需要針對新的關鍵字被索引，則該設備可在提取檔案中逐步地讀取(或同時一或多個提取檔案，或一次一個無損縮減資料塊)、重建原始檔案(但不干擾無損縮減儲存資料)，並解析重建的檔案以更新倒置索引。所有這一切的同時，整個資料庫可以繼續保持以無損縮減形式儲存。

【0268】第16A圖顯示第14A圖所示的模式的變化的模式宣告。第16A圖的模式包含次要維度1609與關鍵字1610的列表的宣告。第16B圖顯示具有在結構描述1602中描述的結構的輸入資料集1611的範例，其被解析並轉換成一組具有根據所宣告的主要維度的名稱的候選元件。該候選元件被轉換成提取之檔案1613中的元件。次要維度“PROD_ID”的宣告對於衍生器施加約束，使得候選元件58可能無法從主要資料元件“NIKESHOELAHJUN with PROD_ID=348”衍生，因此，一個額外的主要資料元件“NIKESHOELAHJUN with PROD_ID=349”在主要資料篩中建立。儘管輸入資料集是如同第14B圖所示的，該提取的

結果是得到7個提取元件，但6個主要資料元件。第16C圖顯示作為提取處理的結果而建立的提取檔案、主樹和主要資料元件。

【0269】第16D圖顯示針對次要維度“PROD_ID”所建立的輔助樹。以特定的PROD_ID值遍歷此樹導致具有那特定的PROD_ID的主要資料元件。例如查詢FetchPDEs(維度5，251)，或可選擇地查詢FetchPDEs(PROD_ID，251)，其要求具有PROD_ID=251的主要資料元件，得到主要資料元件WILSBALLLAHNOV。

【0270】第16E圖顯示針對第16A圖的結構1610中宣告的3個關鍵字所建立的倒置索引(針對關鍵字1631的標記倒置索引)。這些關鍵字是FEDERER、LAVER和SHARAPOVA。倒置索引在解析和消費輸入資料集161之後更新。查詢FetchDistilledElements(關鍵字，Federer)將利用倒置索引(而不是主樹或輔助樹)返回元件2、元件3和元件58。

【0271】第17圖顯示作為強化內容關聯資料檢索的整個設備的方塊圖。內容關聯資料檢索引擎1701提供具有模式1704的資料提取設備或包含資料的維度的結構定義。它也提供了具有關鍵字列表1705的設備。它發出用於從提取設備搜索和檢索資料的查詢1702，以及接收查詢的結果作為結果1703。當建立衍生物時，衍生器110被強化以意識到維度的宣告以在維度的位置禁止內容的修改。請注意，來自葉節點資料結構中的條目對於提取檔案中的元件的反

向參照被儲存在主要資料節106中的葉節點資料結構中。同樣地，輔助索引也被儲存在主要資料節106中。也顯示的是，當元件被寫入到提取資料時，藉由衍生器110連同反向參照1709更新的倒置索引1707。此內容關聯資料檢索引擎與其它應用程式(諸如分析、資料倉儲和資料分析應用程式)互動，為他們提供執行查詢的結果。

【0272】綜上所述，強化的資料提取設備致使對於以無損縮減形式儲存的資料的強大的多維度內容關聯的搜索和檢索。

【0273】資料提取(Data Distillation™)設備可用於音頻和視頻資料的無損縮減的目的。由所述方法所完成的資料縮減係藉由從駐存在內容關聯節中的主要資料元件衍生音頻和視頻的組件來實施。現在將描述用於這種目的的方法的應用程式。

【0274】第18A至18B圖顯示用於根據MPEG 1、Layer 3標準(也稱為MP3)的音頻資料壓縮和解壓縮的編碼器和解碼器的方塊圖。MP3是一種數位音頻的音頻編碼形式，其使用有損和無損資料縮減技術的組合來壓縮輸入的聲音。其管理以將壓縮光碟(CD)音頻從1.4Mbps向下壓縮到128Kbps。MP3利用人耳的限制來抑制將不會被大多數人的人耳察覺到的音頻的部分。為了達成這一目標，採用了統稱為感知編碼技術的一組技術，其有損地但不知不覺地縮減音頻資料的片段的大小。感知編碼技術為有損的，並且在這些步驟期間，丟失的資訊不能被復原。這些感知

編碼技術係輔以霍夫曼編碼(Huffman Coding)，即本文先前所述的無損資料縮減技術。

【0275】在MP3中，輸入的音頻流被壓縮成數個小的資料訊框的序列，其每一個都包含訊框標頭和壓縮音頻資料。原始音頻流被週期性地取樣以產生音頻的片段的序列，其接著採用感知編碼和霍夫曼編碼被壓縮，以產生MP3資料訊框的序列。感知編碼和霍夫曼編碼技術都是在音頻資料的每個片段內局部施加。霍夫曼編碼技術局部利用音頻片段內的冗餘，而不是全域地橫跨音頻流。因此，MP3技術並未全域地利用冗餘-亦未橫跨單一音頻流，也未橫跨多個音頻流之間。這代表了超越MP3所可實現技術的進一步資料縮減的機會。

【0276】每一個MP3資料訊框代表26ms的音頻片段。每個訊框儲存1152個取樣並且被細分為各含有576個取樣的兩個區組。如在第18A圖中的編碼器方塊圖中所示的，在數位音頻訊號的編碼期間，採取了時域樣本並經由過濾的處理和藉由修改的離散餘弦轉換(MDCT)轉換成576個頻域取樣。感知編碼技術被用於縮減包含在取樣中的資訊量。感知編碼器的輸出是包含每個頻率線縮減資訊的非均勻量化區組 1810。接著霍夫曼編碼被用來進一步縮減區組的大小。每個區組的576個頻率線可使用多個霍夫曼表對其編碼。霍夫曼編碼的輸出是包含比例因子、霍夫曼編碼位元以及輔助資料的訊框的主要資料組件。側資訊(用於表徵和定位各種欄位)被置於MP3標頭中。該編碼的輸

出是MP3編碼的音頻訊號。在128Kbps的位元率中，MP3訊框的大小為417個或418個位元組。

【0277】第18C圖顯示在第1A圖首先顯示的資料提取設備如何可以被強化，以對於MP3資料執行資料縮減。在第18C圖中顯示的方法將MP3資料分解成候選元件，並且利用細於該元件本身粒度的元件之間的冗餘。對於MP3資料，該區組被選擇作為元件。在一個實施例中，非均勻量化區組1810(如圖18A中顯示)可以被視為元件。在另一個實施例中，元件可以由量化頻率線1854和比例因子1855的串聯組成。

【0278】在第18C圖中，MP3編碼資料1862的流係由資料提取設備1863所接收，且縮減成以無損縮減形式儲存之提取MP3資料1868的流。MP3編碼資料1862的輸入流由數對的MP3標頭和MP3資料的序列所組成。MP3資料包含CRC、邊資訊(Side Information)、主要資料和輔助資料。由所述設備建立的輸出提取MP3資料由類似的序列對所組成(每對係為DistMP3標頭與隨後的無損縮減形式的元件規範)。所述DistMP3標頭含有有別於主要資料的原始訊框的所有元件，即它包含了MP3標頭、CRC、邊資訊和輔助資料。在此提取MP3資料中的元件欄位包含指定為無損縮減形式的區組。解析器/分解器1864執行輸入MP3編碼流的第一解碼(包含執行霍夫曼解碼)以提取量化的頻率線1851與比例因子1852(其顯示在第18B圖中)，並用以產生音頻區組1865作為候選元件。由解析器/分解器所執行的第一解

碼步驟與第18B圖中的同步和錯誤檢查1851、霍夫曼解碼1852與比例因子解碼1853的步驟是相同的步驟--這些步驟係在任何標準MP3解碼器中執行，且在現有技術中是眾所皆知的。主要資料節1866包含作為主要資料元件的區組，其組織成用內容關聯的方式來進行存取。在將區組安裝成主要資料節期間，區組的內容被用來查明區組應安裝在該節中的何處，並用來更新在節的適當葉節點中的骨架資料結構和元資料。隨後，將區組進行霍夫曼編碼和壓縮，使得它可以被儲存在具有不大於駐存在MP3資料中時它佔據的足跡之足跡的節中。每當在節中的區組需要作為衍生器的主要資料元件，在區組提供給衍生器之前，該區組被解壓縮。使用資料提取設備，輸入的音頻區組係由衍生器1870從駐存在節中的主要資料元件(其也是音頻區組)衍生，並且該區組的無損縮減表示或提取表示被建立且放置在提取MP3資料1868之中。這種區組的提取表示被放置在取代原來存在於MP3訊框的主要資料欄位中的霍夫曼編碼資訊的元件欄位。每個元件或區組的提取表示係利用第1H圖所示的形式來編碼--提取資料中的每個元件為主要資料元件(伴隨著對於節中的主要資料元件或主要區組的參照)，或衍生物元件(伴隨對於節中的主要資料元件或主要區組的參照，加上從主要資料元件所參照的來產生衍生物元件的重建程式)。在衍生步驟期間，用於接受衍生物的臨限值可被設定為駐存在被縮減的訊框的主要資料欄位中的原始霍夫曼編碼資訊的大小的一部分。因此，除非該重

建程式和對於主要資料元件的參照的總和小於MP3編碼訊框的對應主要資料欄位的大小的這個部分(即包含的霍夫曼編碼資料)，衍生物不會被接受。如果該重建程式和對於主要資料元件的參照的總和小於編碼的MP3訊框的現有主要資料欄位的大小的這個部分(即包含的霍夫曼編碼資料)，則可以決定接受該衍生物。

【0279】上述的方法致使在全域範圍(橫跨儲存在該設備的多個音頻區組)的冗餘的開發。MP3編碼資料檔案可以被轉變成提取MP3資料，並以無損縮減形式儲存。需要進行檢索時，(採用檢索器1871和重建器1872的)資料檢索處理可以被呼叫來重建MP3編碼資料1873。在第18C圖中所示的設備中，重建器是負責執行重建程式以產生期望的區組。其額外地強化執行所需要的霍夫曼編碼步驟(在第18A圖中顯示為霍夫曼編碼1811)以產生MP3編碼資料。此資料接著可被饋送到標準的MP3解碼器來播放音頻。

【0280】以這種方式，資料提取設備可適配於並用於進一步將MP3音頻檔案的大小縮減。

【0281】在所描述的方案的另一個變化例中，在接收MP3編碼流時，解析器/分解器將整個主要資料欄位作為供衍生的候選元件或作為供安裝到主要資料篩的主要資料元件。在這種變化例中，所有元件將繼續保持霍夫曼編碼，且重建程式將對於那些已經霍夫曼編碼的元件進行操作。資料提取設備的這種變化也可以被用來進一步縮減MP3音頻檔案的大小。

【0282】以類似於前面部分所述與在圖18A-C所顯示的方式，可以使用資料提取(Data Distillation™)設備以供視頻資料的無損縮減。藉由該方法完成的資料縮減是藉由從駐存在內容關聯節中的主要資料元件衍生視頻資料的組件來實施的。視頻資料流由音頻和動畫元件組件組成。已經描述了用於提取音頻組件的方法。現在將處理動畫組件。動畫組件通常被組織為一連串的影像組。一組圖片以I訊框開始，並且通常後面跟著一些預測訊框(稱為P訊框和B訊框)。I訊框通常較龐大並且包含圖片的完整快照，而預測訊框是在採用諸如關於I訊框的運動估計或關於其它衍生訊框的技術之後衍生的。資料提取(Data Distillation™)設備的一些實施例從視頻資料中提取I訊框作為元件並對它們執行資料提取處理，從而將某些I訊框保留為駐存在內容關聯節中的主要資料元件，而其餘I訊框則從主要資料元件中衍生出來。所描述的方法致使了在全域範圍利用冗餘，橫跨在視頻檔案內且橫跨多個視頻檔案的多個I訊框。由於I訊框通常是動畫資料的龐大組件，因此此方法將使動畫組件的足跡縮減。將提取技術應用於音頻組件以及動畫組件兩者將有助於無損地減少視頻資料的整體大小。

【0283】第19圖顯示首先在圖1A中顯示的資料提取設備如何可被強化以對視頻資料執行資料縮減。視頻資料流1902由資料提取設備1903接收，並被縮減為以無損縮減形式儲存的提取視頻資料1908。輸入的視頻資料流1902包含

兩個組件：壓縮的動畫資料和壓縮的音頻資料。由設備建立的流出提取視頻資料還包含兩個組件，即壓縮的動畫資料和壓縮的音頻資料；然而，這些組件藉由資料提取設備1903來進一步縮減大小。解析器/分解器1904從視頻資料流1902提取壓縮的動畫資料和壓縮的音頻資料，並且從壓縮的動畫訊框資料提取(包含執行任何所需的霍夫曼解碼)內訊框(I訊框)和預測訊框。I訊框被用作候選元件1905以在主要資料節1906中執行內容關聯查找。由主要資料節1906返回的該組主要資料元件(其也是I訊框)由衍生器1910用來產生I訊框的無損縮減表示或提取表示，並且無損縮減的I訊框係配置在提取視頻資料1908中。提取表示係使用圖1H中所示的格式來編碼--提取資料中的每個元件是主要資料元件(伴隨著對節中的主要資料元件的參照)或者衍生物元件(伴隨著對節中的主要資料元件的參照，再加上從正被參照的主要資料元件產生衍生物元件的重建程式)。在衍生步驟期間，接受衍生的臨限值可以被設置為原始I訊框的大小的一部分。因此，除非重建程式和對主要資料元件的參照的總和小於對應I訊框的大小的這一部分，否則將不會接受該衍生。如果重建程式和對主要資料元件的參照的總和小於原始I訊框的大小的這一部分，則可以做出接受該衍生物的決定。

【0284】以上所描述的方法致使了在全域範圍利用冗餘，橫跨儲存在設備內的多個視頻資料集的多個I訊框。當需要被檢索時，可以調用(採用檢索器1911和重建器

1912的)資料檢索處理來重建視頻資料1913。在第19圖所示的設備中，重建器負責執行重建程式以產生希望的I訊框。額外地強化以將壓縮的音頻資料與壓縮的動畫資料相結合(已由解析器&分解器1904執行的提取操作的實質倒置)以產生視頻資料1913。此資料可以隨後被饋送到標準視頻解碼器以播放視頻。

【0285】以這種方式，資料提取設備可適配於並用於進一步將視頻檔案的大小縮減。

【0286】上述說明係提出以使熟習本領域之任何人士能做成及使用該等實施例。對所揭示之實施例的各種修正對於熟習本領域的技術人員將是顯而易見的，且在此所界定的一般原則可應用至其它的實施例及應用，而不會背離本發明之精神和範圍。因此，本發明不應受限於所顯示之實施例，而是應根據與在此所揭示的原理及特徵一致的最廣泛範圍。

【0287】在此發明中所敘述的資料結構及代碼可被部分地，或完全地儲存在電腦可讀取儲存媒體及/或硬體模組及/或硬體設備上。電腦可讀取儲存媒體包含(但不受限於)揮發性記憶體、非揮發性記憶體、諸如碟片驅動器之磁性及光學儲存裝置、磁帶、CD(光碟)、DVD(數位多功能碟片或數位視頻碟片)、或目前所已知或爾後將發展之能夠儲存代碼及/或資料的其它媒體。在此發明中所敘述之硬體模組或設備包含(但不受限於)特殊應用積體電路(ASIC)、現場可程式化閘陣列(FPGA)、專用或共享之處理

器、及/或目前所已知或爾後將發展之其它的硬體模組或設備。

【0288】在此發明中所敘述之方法及處理可被部分地或完全地實施為被儲存在電腦可讀取儲存媒體或裝置中的代碼及/或資料，使得當電腦系統讀取及執行所述代碼及/或資料時，該電腦系統可執行相關聯的方法及處理。所述方法及處理亦可被部分地，或完全地實施於硬體模組或設備中，使得當所述硬體模組或設備被致動時，它們可執行相關聯的方法及處理。應注意的是，所述方法及處理可使用代碼、資料及硬體模組或設備的組合來實施。

【0289】本發明之實施例的上述說明僅被呈現以供描繪及說明的目的之用。它們並不打算要變成包羅無遺的，或要限制本發明至所揭示的形式。因而，許多修正和變化將對於熟習本領域之從業者是顯而易見的。此外，以上的揭示並不打算要限制本發明。

【符號說明】

【0290】

102：輸入資料

103：資料縮減設備

104：解析器及分解器

105：候選元件

106：資料提取篩或主要資料篩

107,130,305-307：主要資料元件

- 108：提取之資料
- 109：檢索請求
- 110：衍生器
- 111：檢索器
- 112：重建器
- 113：檢索資料輸出
- 114：更新
- 115：縮減資料組件
- 116：重建之資料
- 119A：重建程式
- 119B：參照或指標
- 121, 122：內容相關映射器
- 131：參照
- 132：主要重建程式
- 133：參照
- 202-222：操作
- 301：路徑
- 302：根
- 303：節點
- 334,336,340,354：鏈路
- 1201：輸入檔案
- 1203：資料提取設備
- 1205：提取之檔案
- 1206：主要資料節或主要資料儲存

- 1207：映射器
- 1209：樹狀節點檔案
- 1210：葉狀節點檔案
- 1211：PDE檔案
- 1212：檔案1
- 1213：檔案1.提取
- 1221：輸入資料集
- 1224：輸入資料批量
- 1226：資料批量i
- 1227：無損縮減資料集
- 1228：無損縮減資料批量i
- 1251：輸入檔案
- 1252：候選元件
- 1253：無損縮減表示
- 1254：無損縮減表示
- 1270：從屬提取模組1
- 1271：輸入檔案I
- 1275：候選元件
- 1276：無損縮減元件
- 1277：候選元件
- 1278：無損縮減元件
- 1279：從屬提取模組2
- 1280：包裹
- 1281：標頭

- 1282：包裹長度
- 1283：提取之檔案
- 1284：PDE檔案
- 1285：來源清單
- 1286：目的地清單和映射器
- 1402：結構描述
- 1404：維度映射描述
- 1405：檔案
- 1406：候選元件
- 1407：名稱
- 1408：提取之檔案
- 1602：結構描述
- 1609：次要維度
- 1610：關鍵字
- 1611：輸入資料集
- 1613：提取之檔案
- 1631：關鍵字
- 1701：內容關聯資料檢索引擎
- 1702：查詢
- 1703：結果
- 1704：模式
- 1705：關鍵字列表
- 1707：倒置索引
- 1709：反向參照

- 1810：非均勻量化區組
- 1811：霍夫曼編碼
- 1851：同步和錯誤檢查
- 1852：霍夫曼解碼
- 1853：比例因子解碼
- 1854：量化頻率線
- 1855：比例因子
- 1862：MP3編碼資料
- 1863：資料提取設備
- 1864：解析器/分解器
- 1865：音頻區組
- 1866：主要資料節
- 1868：提取MP3資料
- 1870：衍生器
- 1871：檢索器
- 1872：重建器
- 1873：MP3編碼資料
- 1902：視頻資料流
- 1903：資料提取設備
- 1904：解析器/分解器
- 1905：候選元件
- 1906：主要資料節
- 1908：提取視頻資料
- 1910：衍生器

- 1911：檢 索 器
- 1912：重 建 器
- 1913：視 頻 資 料

【發明申請專利範圍】

【第 1 項】

一種用於將視頻資料縮減以獲得縮減視頻資料的方法，該方法包含：

將來自該視頻資料的壓縮動畫資料和壓縮音頻資料解壓縮；

將來自該壓縮動畫資料的內訊框(I訊框)解壓縮；

將來自該壓縮動畫資料的預測訊框解壓縮；

維持組織主要資料元件的資料結構，每個主要資料元件包含位元組的序列；

無損縮減該 I 訊框以獲得無損縮減的 I 訊框，其中對於每個 I 訊框，無損縮減該 I 訊框包含：

藉由使用該 I 訊框來識別第一組主要資料元件，以對於根據其內容組織主要資料元件的資料結構執行第一內容關聯查找，以及

使用該第一組主要資料元件來無損縮減該 I 訊框，以 (i) 在該 I 訊框為該第一組主要資料元件中主要資料元件的副本的情況下，獲得對該主要資料元件的參照；或者 (ii) 在該 I 訊框並非該第一組主要資料元件中任何主要資料元件的副本的情況下，獲得對該第一組主要資料元件中一或多個主要資料元件的參照，以及獲得從該一或多個主要資料元件得到該 I 訊框的轉換的序列；以及

其中，該縮減視頻資料包含該無損縮減的 I 訊

框、該無損縮減的 I 訊框所參照的主要資料元件、該預測訊框和該壓縮音頻資料。

【第 2 項】

如申請專利範圍第 1 項的方法，其中使用該第一組主要資料元件來無損縮減該 I 訊框包含：

響應於確定 (i) 對於該第一組主要資料元件之參照的大小和 (ii) 重建程式的描述的大小之總和係小於該 I 訊框的大小的臨限值部分，產生該 I 訊框的第一無損縮減表示，其中該第一無損縮減表示包含對於在該第一組主要資料元件中的每個主要資料元件的參照和該重建程式的描述；以及

響應於確定 (i) 對於該第一組主要資料元件之該參照的該大小和 (ii) 該重建程式的該描述的該大小之該總和係大於或等於該 I 訊框的該大小的該臨限值部分，

在該資料結構中加入該 I 訊框作為新的主要資料元件，以及

產生該 I 訊框的第二無損縮減表示，其中該第二無損縮減表示包含對於該新的主要資料元件的參照。

【第 3 項】

如申請專利範圍第 2 項的方法，其中該重建程式的該描述指定一連串的轉換，其當施加到該第一組主要資料元件時，產生該 I 訊框。

【第 4 項】

如申請專利範圍第 1 項的方法，其中該方法還包含：

將該壓縮的音頻資料解壓縮，以獲得一組音頻組件；

以及

對於該組音頻組件中的每一個音頻組件，

藉由使用該音頻組件來識別第二組主要資料元件，以對於根據其內容組織主要資料元件的該資料結構執行第二內容關聯查找，以及

使用該第二組主要資料元件來無損縮減該音頻組件；以及

其中，該縮減視頻資料包含該無損縮減的I訊框、該無損縮減的I訊框所參照的主要資料元件、該預測訊框和無損縮減的音頻組件和該無損縮減的音頻組件所參照的主要資料元件。

【第5項】

如申請專利範圍第4項的方法，其中使用該第二組主要資料元件來無損縮減該音頻組件包含：

響應於確定(i)對於該第二組主要資料元件之參照的大小和(ii)重建程式的描述的大小之總和係小於該音頻組件的大小的臨限值部分，產生該音頻組件的第一無損縮減表示，其中該第一無損縮減表示包含對於在該第二組主要資料元件中的每個主要資料元件的參照和該重建程式的描述；以及

響應於確定(i)對於該第二組主要資料元件之該參照的該大小和(ii)該重建程式的該描述的該大小之該總和係大於或等於該音頻組件的該大小的該臨限值部分，

在該資料結構中加入該音頻組件作為新的主要資料元

件，以及

產生該音頻組件的第二無損縮減表示，其中該第二無損縮減表示包含對於該新的主要資料元件的參照。

【第6項】

如申請專利範圍第5項的方法，其中該重建程式的該描述指定一連串的轉換，其當施加到該第二組主要資料元件時，產生該音頻組件。

【第7項】

如申請專利範圍第1項的方法，還包含：

響應於收到檢索該視頻資料的請求，

重建該壓縮動畫資料，其包含從該無損縮減的I訊框和該無損縮減的I訊框所參照的該主要資料元件重新構建該I訊框，並將該I訊框與該預測訊框結合，以及

將該壓縮動畫資料與該壓縮音頻資料結合以獲得該視頻資料。

【第8項】

如申請專利範圍第1項的方法，其中所述從該壓縮動畫資料中提取該I訊框包括執行霍夫曼解碼。

【第9項】

如申請專利範圍第8項的方法，還包含對該無損縮減的I訊框和該無損縮減的I訊框所參照的主要資料元件執行霍夫曼編碼，以分別獲得霍夫曼編碼的無損縮減的I訊框和霍夫曼編碼的主要資料元件，其中該縮減視頻資料包含該霍夫曼編碼的無損縮減的I訊框、該霍夫曼編碼的主要

資料元件、該預測訊框和該壓縮音頻資料。

【第10項】

如申請專利範圍第9項的方法，還包含：

響應於收到檢索該視頻資料的請求，

透過(1)對該霍夫曼編碼的無損縮減的I訊框進行霍夫曼解碼，以獲得該無損縮減的I訊框，並且對該霍夫曼編碼的主要資料元件進行霍夫曼解碼，以獲得該主要資料元件、(2)從該無損縮減的I訊框和該主要資料元件重新構建該I訊框，和(3)對該I訊框執行霍夫曼編碼並與該預測訊框結合，以重建該壓縮動畫資料，以及

將該壓縮動畫資料與該壓縮音頻資料結合以獲得該視頻資料。

【第11項】

一種電腦可讀取儲存媒體，其儲存指令，當所述指令由電腦執行時使該電腦執行用於將視頻資料縮減以獲得縮減視頻資料的方法，該方法包含：

將來自該視頻資料的壓縮動畫資料和壓縮音頻資料解壓縮；

將來自該壓縮動畫資料的內訊框(I訊框)解壓縮；

將來自該壓縮動畫資料的預測訊框解壓縮；

維持組織主要資料元件的資料結構，每個主要資料元件包含位元組的序列；

無損縮減該I訊框以獲得無損縮減的I訊框，其中對於每個I訊框，無損縮減該I訊框包含：

藉由使用該 I 訊框來識別第一組主要資料元件，以對於根據其內容組織主要資料元件的資料結構執行第一內容關聯查找，以及

使用該第一組主要資料元件來無損縮減該 I 訊框，以 (i) 在該 I 訊框為該第一組主要資料元件中主要資料元件的副本的情況下，獲得對該主要資料元件的參照；或者 (ii) 在該 I 訊框並非該第一組主要資料元件中任何主要資料元件的副本的情況下，獲得對該第一組主要資料元件中一或多個主要資料元件的參照，以及獲得從該一或多個主要資料元件得到該 I 訊框的轉換的序列；以及

其中，該縮減視頻資料包含該無損縮減的 I 訊框、該無損縮減的 I 訊框所參照的主要資料元件、該預測訊框和該壓縮音頻資料。

【第 12 項】

如申請專利範圍第 11 項的電腦可讀取儲存媒體，其中使用該第一組主要資料元件來無損縮減該 I 訊框包含：

響應於確定 (i) 對於該第一組主要資料元件之參照的大小和 (ii) 重建程式的描述的大小之總和係小於該 I 訊框的大小的臨限值部分，產生該 I 訊框的第一無損縮減表示，其中該第一無損縮減表示包含對於在該第一組主要資料元件中的每個主要資料元件的參照和該重建程式的描述；以及

響應於確定 (i) 對於該第一組主要資料元件之該參照的該大小和 (ii) 該重建程式的該描述的該大小之該總和係大

於或等於該 I 訊框的該大小的該臨限值部分，

在該資料結構中加入該 I 訊框作為新的主要資料元件，以及

產生該 I 訊框的第二無損縮減表示，其中該第二無損縮減表示包含對於該新的主要資料元件的參照。

【第 13 項】

如申請專利範圍第 12 項的電腦可讀取儲存媒體，其中該重建程式的該描述指定一連串的轉換，其當施加到該第一組主要資料元件時，產生該 I 訊框。

【第 14 項】

如申請專利範圍第 11 項的電腦可讀取儲存媒體，其中該方法還包含：

將該壓縮的音頻資料解壓縮，以獲得一組音頻組件；
以及

對於該組音頻組件中的每一個音頻組件，

藉由使用該音頻組件來識別第二組主要資料元件，以對於根據其內容組織主要資料元件的該資料結構執行第二內容關聯查找，以及

使用該第二組主要資料元件來無損縮減該音頻組件；
以及

其中，該縮減視頻資料包含該無損縮減的 I 訊框、該無損縮減的 I 訊框所參照的主要資料元件、該預測訊框和無損縮減的音頻組件和該無損縮減的音頻組件所參照的主要資料元件。

【第15項】

如申請專利範圍第14項的電腦可讀取儲存媒體，其中使用該第二組主要資料元件來無損縮減該音頻組件包含：

響應於確定(i)對於該第二組主要資料元件之參照的大小和(ii)重建程式的描述的大小之總和係小於該音頻組件的大小的臨限值部分，產生該音頻組件的第一無損縮減表示，其中該第一無損縮減表示包含對於在該第二組主要資料元件中的每個主要資料元件的參照和該重建程式的描述；以及

響應於確定(i)對於該第二組主要資料元件之該參照的該大小和(ii)該重建程式的該描述的該大小之該總和係大於或等於該音頻組件的該大小的該臨限值部分，

在該資料結構中加入該音頻組件作為新的主要資料元件，以及

產生該音頻組件的第二無損縮減表示，其中該第二無損縮減表示包含對於該新的主要資料元件的參照。

【第16項】

如申請專利範圍第15項的電腦可讀取儲存媒體，其中該重建程式的該描述指定一連串的轉換，其當施加到該第二組主要資料元件時，產生該音頻組件。

【第17項】

如申請專利範圍第11項的電腦可讀取儲存媒體，還包含：

響應於收到檢索該視頻資料的請求，

重建該壓縮動畫資料，其包含從該無損縮減的 I 訊框和該無損縮減的 I 訊框所參照的該主要資料元件重新構建該 I 訊框，並將該 I 訊框與該預測訊框結合，以及

將該壓縮動畫資料與該壓縮音頻資料結合以獲得該視頻資料。

【第 18 項】

如申請專利範圍第 11 項的電腦可讀取儲存媒體，其中所述從該壓縮動畫資料中提取該 I 訊框包括執行霍夫曼解碼。

【第 19 項】

如申請專利範圍第 18 項的電腦可讀取儲存媒體，還包含對該無損縮減的 I 訊框和該無損縮減的 I 訊框所參照的主要資料元件執行霍夫曼編碼，以分別獲得霍夫曼編碼的無損縮減的 I 訊框和霍夫曼編碼的主要資料元件，其中該縮減視頻資料包含該霍夫曼編碼的無損縮減的 I 訊框、該霍夫曼編碼的主要資料元件、該預測訊框和該壓縮音頻資料。

【第 20 項】

如申請專利範圍第 19 項的電腦可讀取儲存媒體，還包含：

響應於收到檢索該視頻資料的請求，

透過 (1) 對該霍夫曼編碼的無損縮減的 I 訊框進行霍夫曼解碼，以獲得該無損縮減的 I 訊框，並且對該霍夫曼編碼的主要資料元件進行霍夫曼解碼，以獲得該主要資

料元件、(2)從該無損縮減的I訊框和該主要資料元件重新構建該I訊框，和(3)對該I訊框執行霍夫曼編碼並與該預測訊框結合，以重建該壓縮動畫資料，以及

將該壓縮動畫資料與該壓縮音頻資料結合以獲得該視頻資料。

【第21項】

一種電子設備，包含：

處理器；以及

記憶體，其儲存指令，當所述指令由該處理器執行時使該處理器執行用於將視頻資料縮減以獲得縮減視頻資料的方法，該方法包含：

將來自該視頻資料的壓縮動畫資料和壓縮音頻資料解壓縮；

將來自該壓縮動畫資料的內訊框(I訊框)解壓縮；

將來自該壓縮動畫資料的預測訊框解壓縮；

維持組織主要資料元件的資料結構，每個主要資料元件包含位元組的序列；

無損縮減該I訊框以獲得無損縮減的I訊框，其中對於每個I訊框，無損縮減該I訊框包含：

藉由使用該I訊框來識別第一組主要資料元件，以對於根據其內容組織主要資料元件的資料結構執行第一內容關聯查找，以及

使用該第一組主要資料元件來無損縮減該I訊框，以(i)在該I訊框為該第一組主要資料元件中主要

資料元件的副本的情況下，獲得對該主要資料元件的參照；或者(ii)在該I訊框並非該第一組主要資料元件中任何主要資料元件的副本的情況下，獲得對該第一組主要資料元件中一或多個主要資料元件的參照，以及獲得從該一或多個主要資料元件得到該I訊框的轉換的序列；以及

其中，該縮減視頻資料包含該無損縮減的I訊框、該無損縮減的I訊框所參照的主要資料元件、該預測訊框和該壓縮音頻資料。

【第22項】

如申請專利範圍第21項的電子設備，其中使用該第一組主要資料元件來無損縮減該I訊框包含：

響應於確定(i)對於該第一組主要資料元件之參照的大小和(ii)重建程式的描述的大小之總和係小於該I訊框的大小的臨限值部分，產生該I訊框的第一無損縮減表示，其中該第一無損縮減表示包含對於在該第一組主要資料元件中的每個主要資料元件的參照和該重建程式的描述；以及

響應於確定(i)對於該第一組主要資料元件之該參照的該大小和(ii)該重建程式的該描述的該大小之該總和係大於或等於該I訊框的該大小的該臨限值部分，

在該資料結構中加入該I訊框作為新的主要資料元件，以及

產生該I訊框的第二無損縮減表示，其中該第二無損縮減表示包含對於該新的主要資料元件的參照。

【第23項】

如申請專利範圍第22項的電子設備，其中該重建程式的該描述指定一連串的轉換，其當施加到該第一組主要資料元件時，產生該I訊框。

【第24項】

如申請專利範圍第21項的電子設備，其中該方法還包含：

將該壓縮的音頻資料解壓縮，以獲得一組音頻組件；
以及

對於該組音頻組件中的每一個音頻組件，

藉由使用該音頻組件來識別第二組主要資料元件，以對於根據其內容組織主要資料元件的該資料結構執行第二內容關聯查找，以及

使用該第二組主要資料元件來無損縮減該音頻組件；
以及

其中，該縮減視頻資料包含該無損縮減的I訊框、該無損縮減的I訊框所參照的主要資料元件、該預測訊框和無損縮減的音頻組件和該無損縮減的音頻組件所參照的主要資料元件。

【第25項】

如申請專利範圍第24項的電子設備，其中使用該第二組主要資料元件來無損縮減該音頻組件包含：

響應於確定(i)對於該第二組主要資料元件之參照的大小和(ii)重建程式的描述的大小之總和係小於該音頻組件

的大小的臨限值部分，產生該音頻組件的第一無損縮減表示，其中該第一無損縮減表示包含對於在該第二組主要資料元件中的每個主要資料元件的參照和該重建程式的描述；以及

響應於確定(i)對於該第二組主要資料元件之該參照的該大小和(ii)該重建程式的該描述的該大小之該總和係大於或等於該音頻組件的該大小的該臨限值部分，

在該資料結構中加入該音頻組件作為新的主要資料元件，以及

產生該音頻組件的第二無損縮減表示，其中該第二無損縮減表示包含對於該新的主要資料元件的參照。

【第26項】

如申請專利範圍第25項的電子設備，其中該重建程式的該描述指定一連串的轉換，其當施加到該第二組主要資料元件時，產生該音頻組件。

【第27項】

如申請專利範圍第21項的電子設備，還包含：

響應於收到檢索該視頻資料的請求，

重建該壓縮動畫資料，其包含從該無損縮減的I訊框和該無損縮減的I訊框所參照的該主要資料元件重新構建該I訊框，並將該I訊框與該預測訊框結合，以及

將該壓縮動畫資料與該壓縮音頻資料結合以獲得該視頻資料。

【第28項】

如申請專利範圍第21項的電子設備，其中所述從該壓縮動畫資料中提取該I訊框包括執行霍夫曼解碼。

【第29項】

如申請專利範圍第28項的電子設備，還包含對該無損縮減的I訊框和該無損縮減的I訊框所參照的主要資料元件執行霍夫曼編碼，以分別獲得霍夫曼編碼的無損縮減的I訊框和霍夫曼編碼的主要資料元件，其中該縮減視頻資料包含該霍夫曼編碼的無損縮減的I訊框、該霍夫曼編碼的主要資料元件、該預測訊框和該壓縮音頻資料。

【第30項】

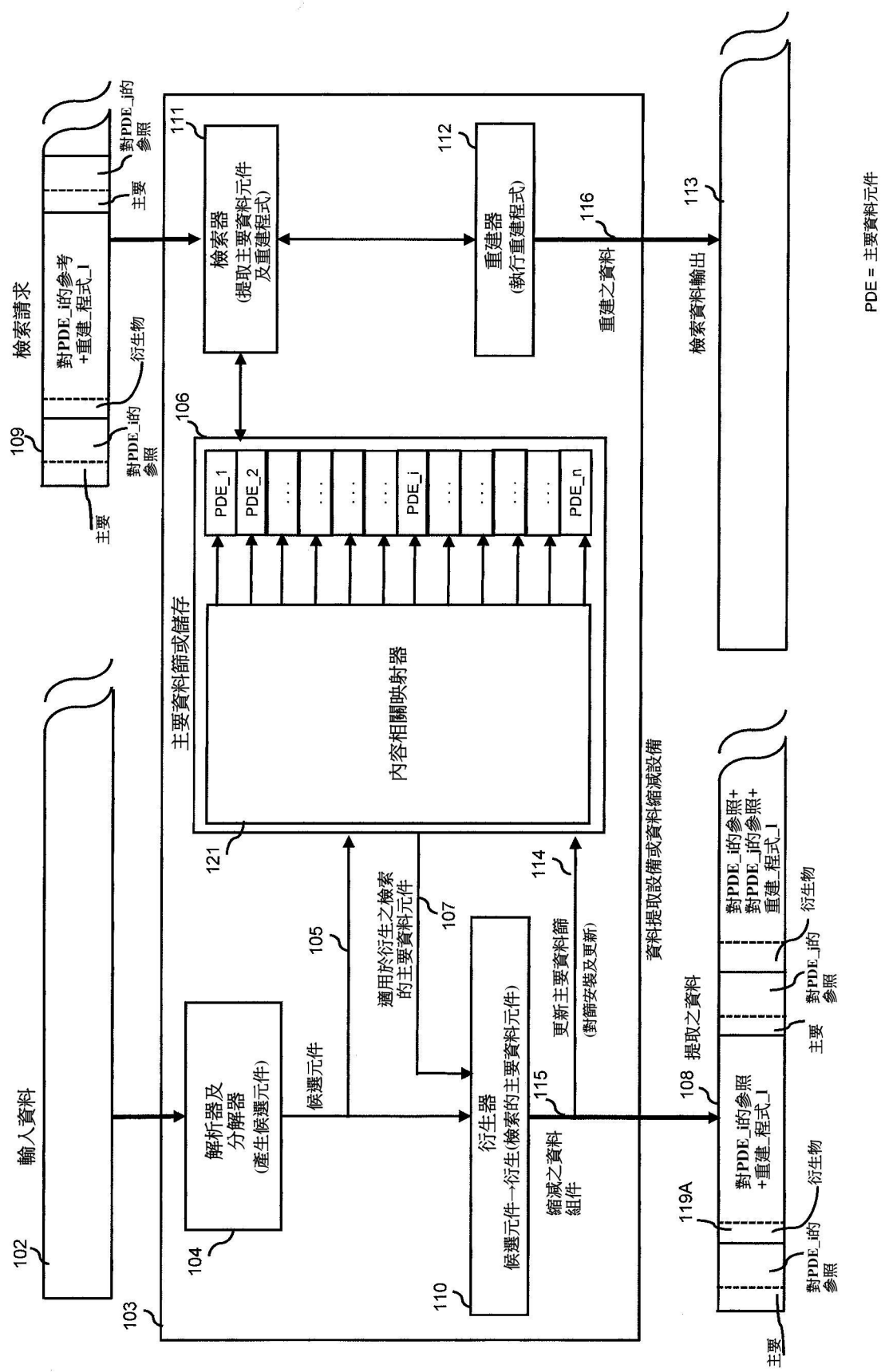
如申請專利範圍第29項的電子設備，還包含：

響應於收到檢索該視頻資料的請求，

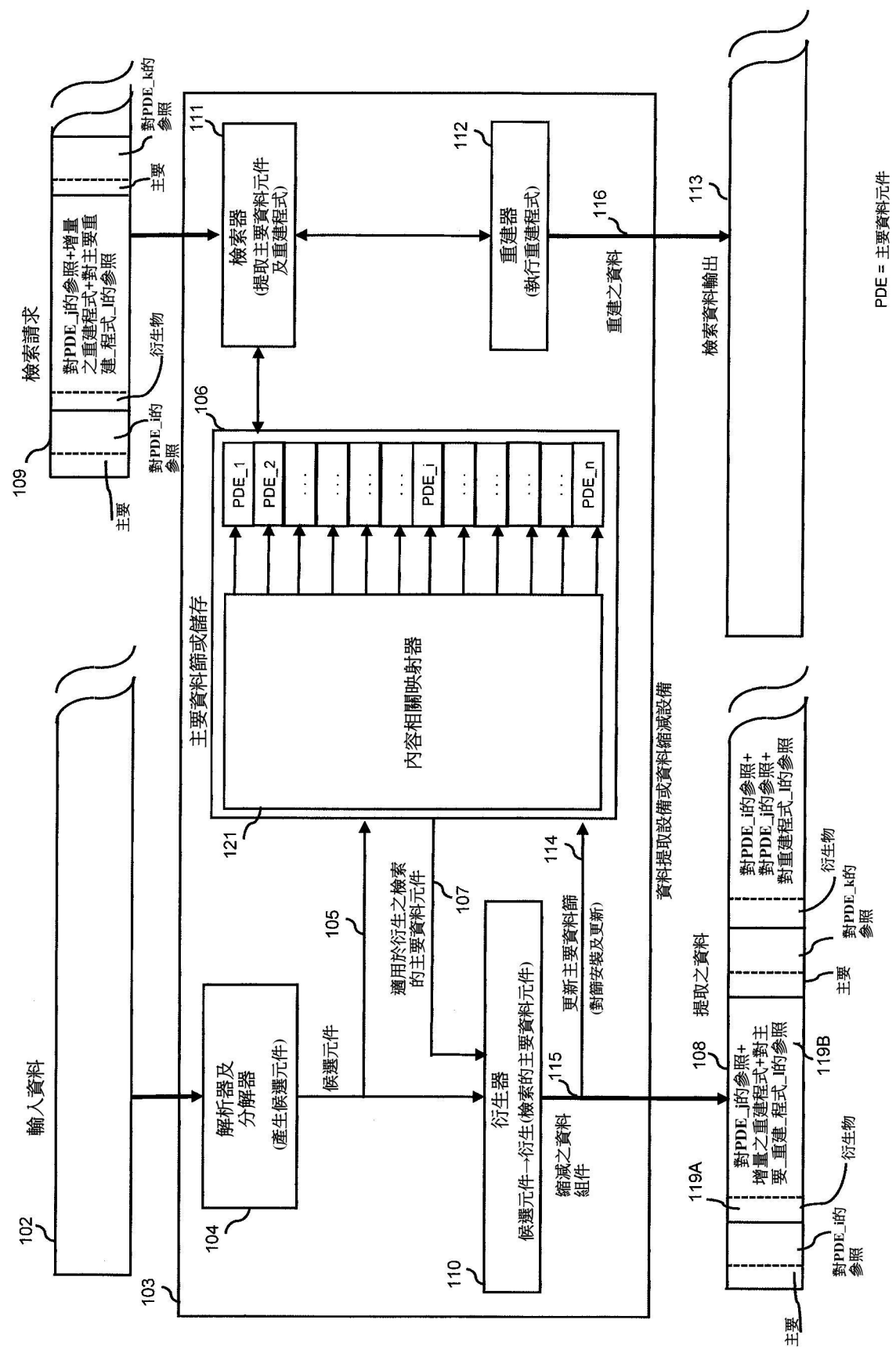
透過(1)對該霍夫曼編碼的無損縮減的I訊框進行霍夫曼解碼，以獲得該無損縮減的I訊框，並且對該霍夫曼編碼的主要資料元件進行霍夫曼解碼，以獲得該主要資料元件、(2)從該無損縮減的I訊框和該主要資料元件重新構建該I訊框，和(3)對該I訊框執行霍夫曼編碼並與該預測訊框結合，以重建該壓縮動畫資料，以及

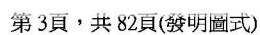
將該壓縮動畫資料與該壓縮音頻資料結合以獲得該視頻資料。

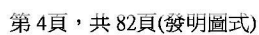
第1A圖

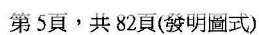


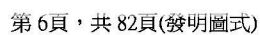
第1B圖

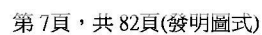












第1H圖

欄名	尺寸	說明	編碼
記錄類型	3 位元	指示提取之元件的類型	0: 主要資料元件在主要資料儲存中 1: 具有單一PDE重建的衍生物(RP直列在提取之資料中) 2: 具有單一PDE重建的衍生物(PRP在主要資料儲存中) 3: 具有單一PDE重建的衍生物(PRP在PRP儲存中) 4: 具有多個PDE重建的衍生物(RP直列在提取之資料中) 5: 具有多個PDE重建的衍生物(PRP在主要資料儲存中) 6: 具有多個PDE重建的衍生物(PRP在PRP儲存中) 7: 主要資料元件直列(在提取之資料中) 注意: PRP=重建程式, PDE=主要資料元件
元件尺寸	2 位元組	元件的尺寸 (在原始輸入中之元件的尺寸)	編碼為(尺寸-1) × 0=1位元組, 65535=64K位元組
額外PDE 計數	2 位元	僅用於多個PDE重建: 額外的主要資料元件之數目	編碼為(計數-1) × 0=1尺寸, 3=4尺寸
參照尺寸	3 位元	對主要資料元件或重建程式之 以位元組的參照之尺寸(RP)	編碼為(尺寸-1) × 0=1位元組, 1=2位元組, 7=8位元組
PDE 尺寸	2 位元組	主要資料元件(PDE)之尺寸	編碼為(尺寸-1) × 0=1位元組, 65535=64K位元組
PDE 參照	(參照尺寸)	用於缺設的主要資料元件及 任何額外的主要資料元件之尺寸的參照	元件參照
RP 尺寸	2 位元組	重建程式之尺寸(RP)	編碼為(尺寸-1) × 0=1位元組, 65535=64K位元組
重建程式 參照	(參照尺寸)	僅用於主要資料儲存或重建程式儲存中之 重建程式: 對重建程式的參照	重建程式參照
重建程式 (RP)	可變的	僅用於重建程式直列: 重建程式	請參閱第7A圖

實例1: 複製品或主要資料元件

記錄類型: 0	元件尺寸: 2位元組	參照尺寸: 5(6位元組)	PDE尺寸: 2位元組	PDE參照: 6位元組
---------	------------	---------------	-------------	-------------

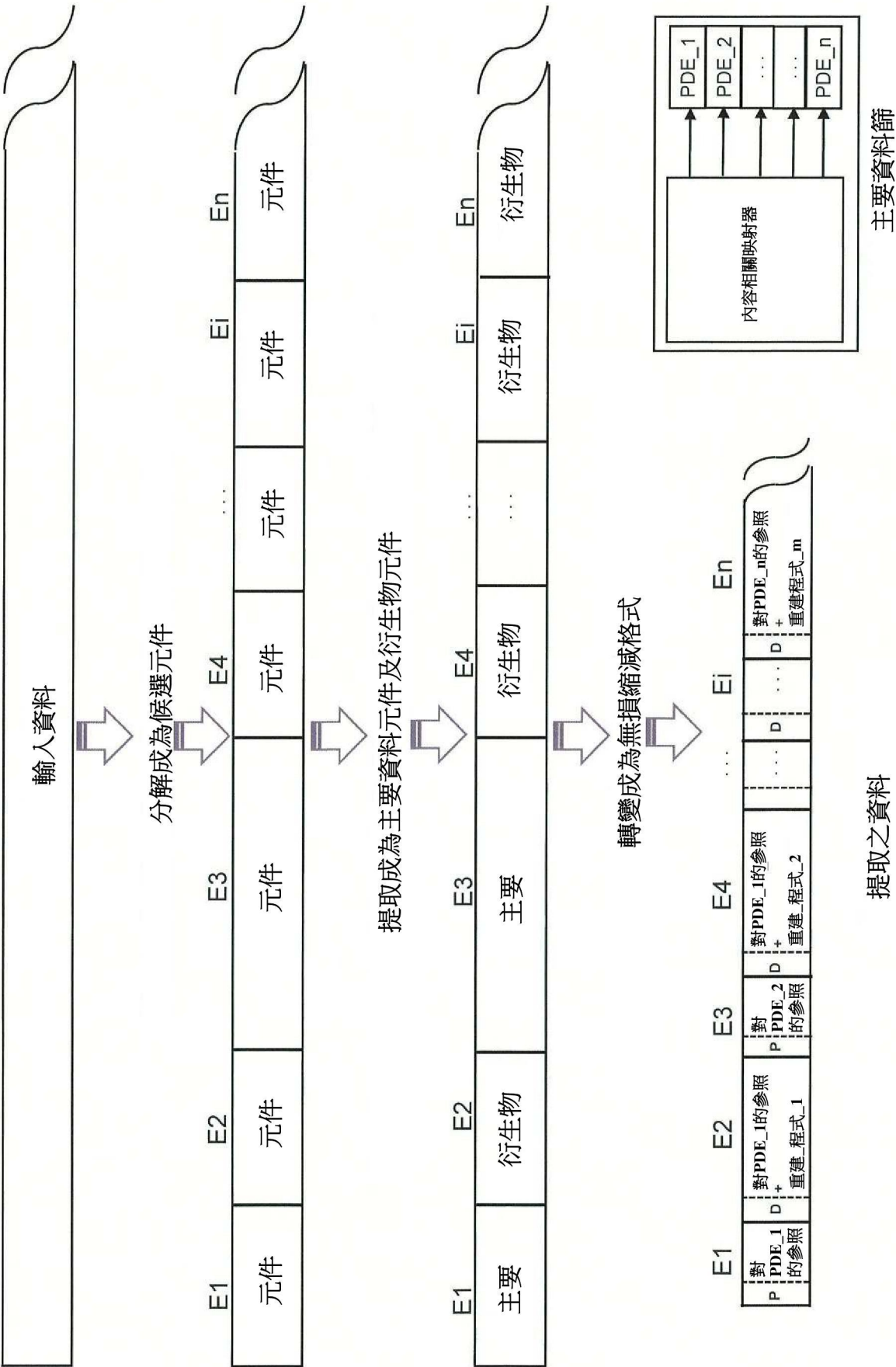
實例2: 具有單一元件重建之衍生物, RP直列(13位元組)

記錄類型: 1	元件尺寸: 2位元組	參照尺寸: 5(6位元組)	PDE尺寸: 2位元組	PDE參照: 6位元組	RP尺寸: 2位元組	重建程式: 13位元組
---------	------------	---------------	-------------	-------------	------------	-------------

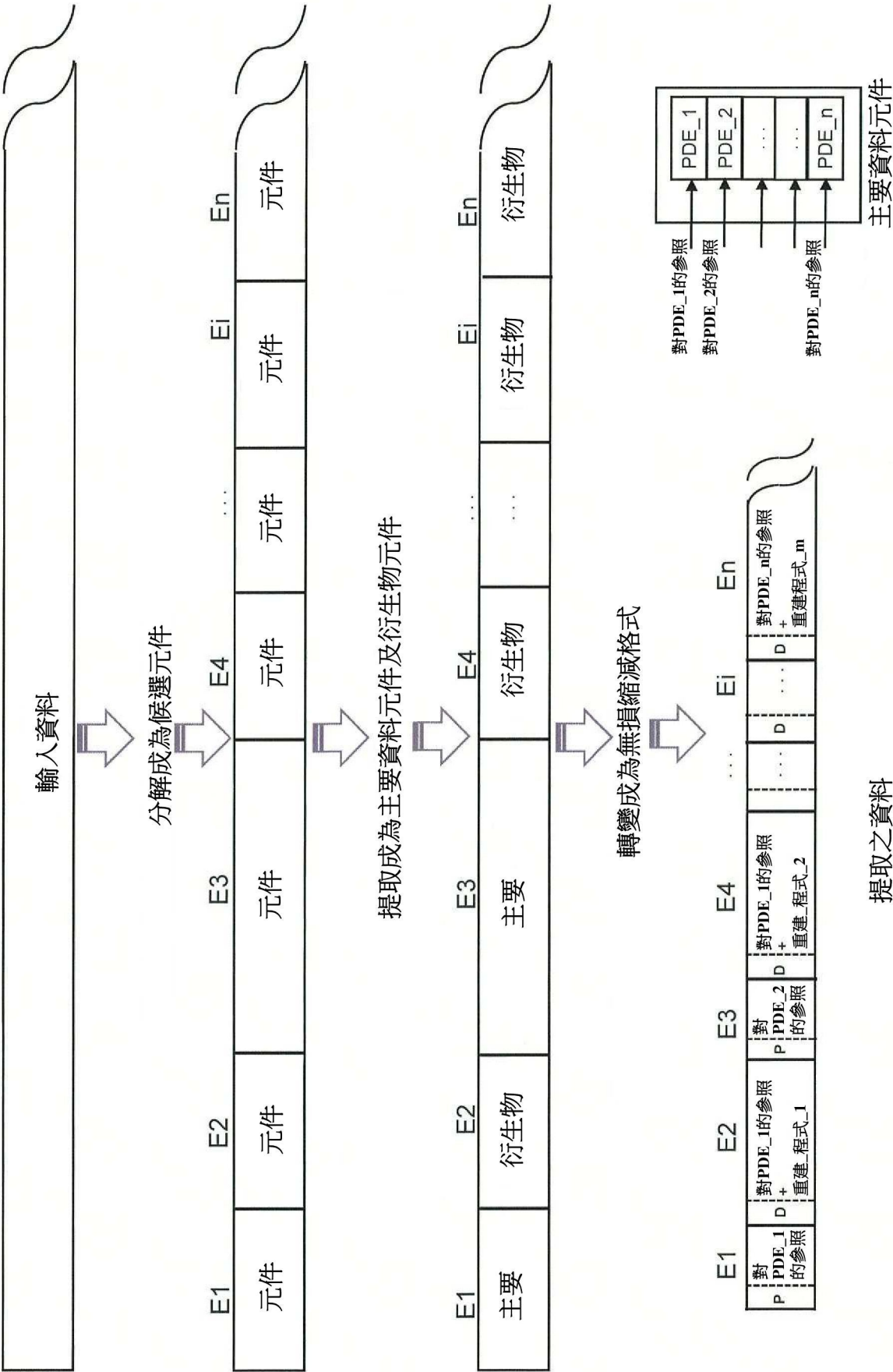
實例3: 具有多個元件重建之衍生物, 2主要資料元件, RP在元件儲存中(注意:B位元組)

記錄類型: 5	元件尺寸: 2B	額外PDE計數: 1	參照尺寸: 5	PDE1尺寸: 2B	PDE1參照: 6B	PDE2尺寸: 2B	PDE2參照: 6B	RP尺寸: 6B	RP參照: 2B
---------	----------	------------	---------	------------	------------	------------	------------	----------	----------

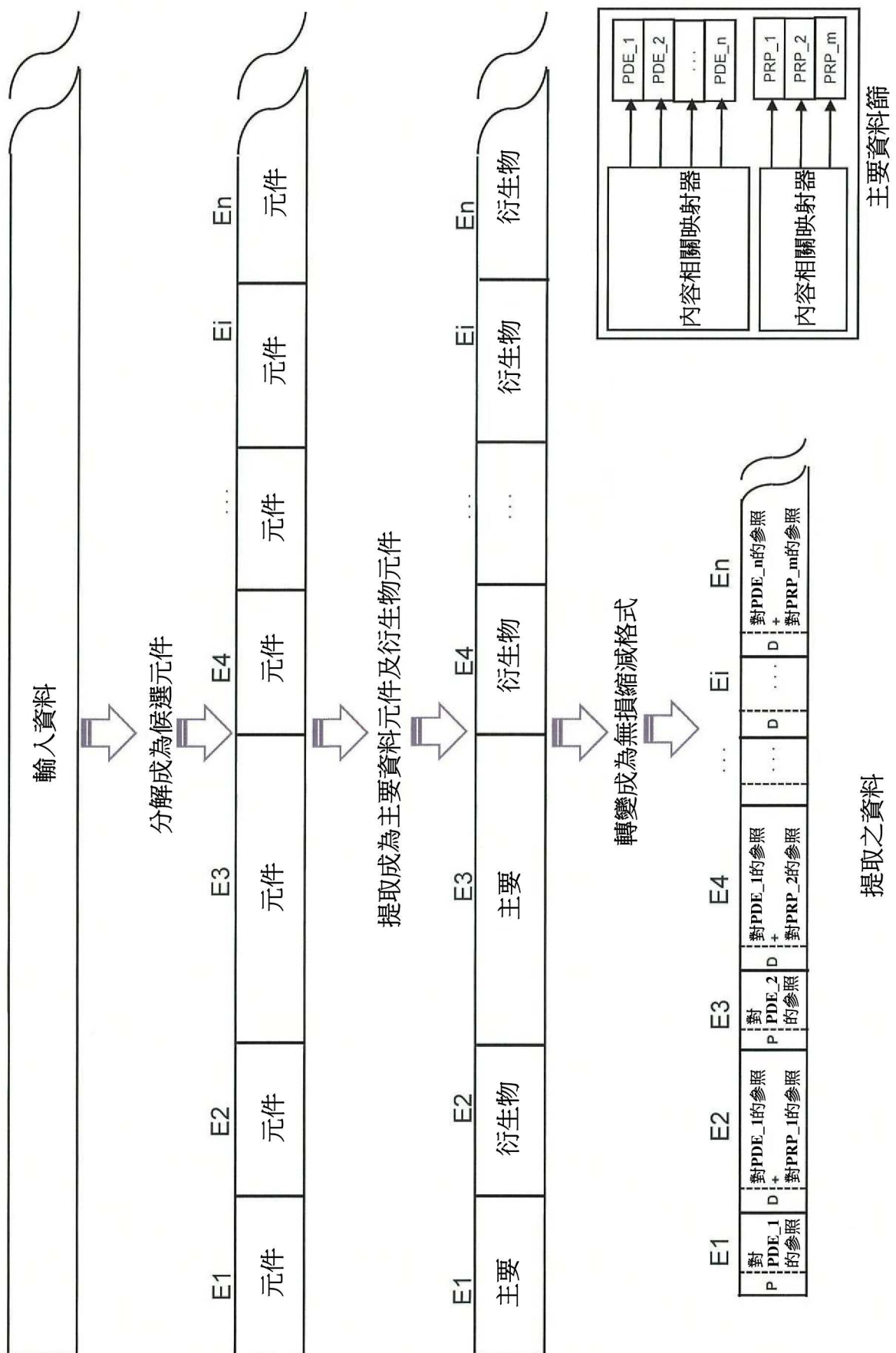
第 1I 圖



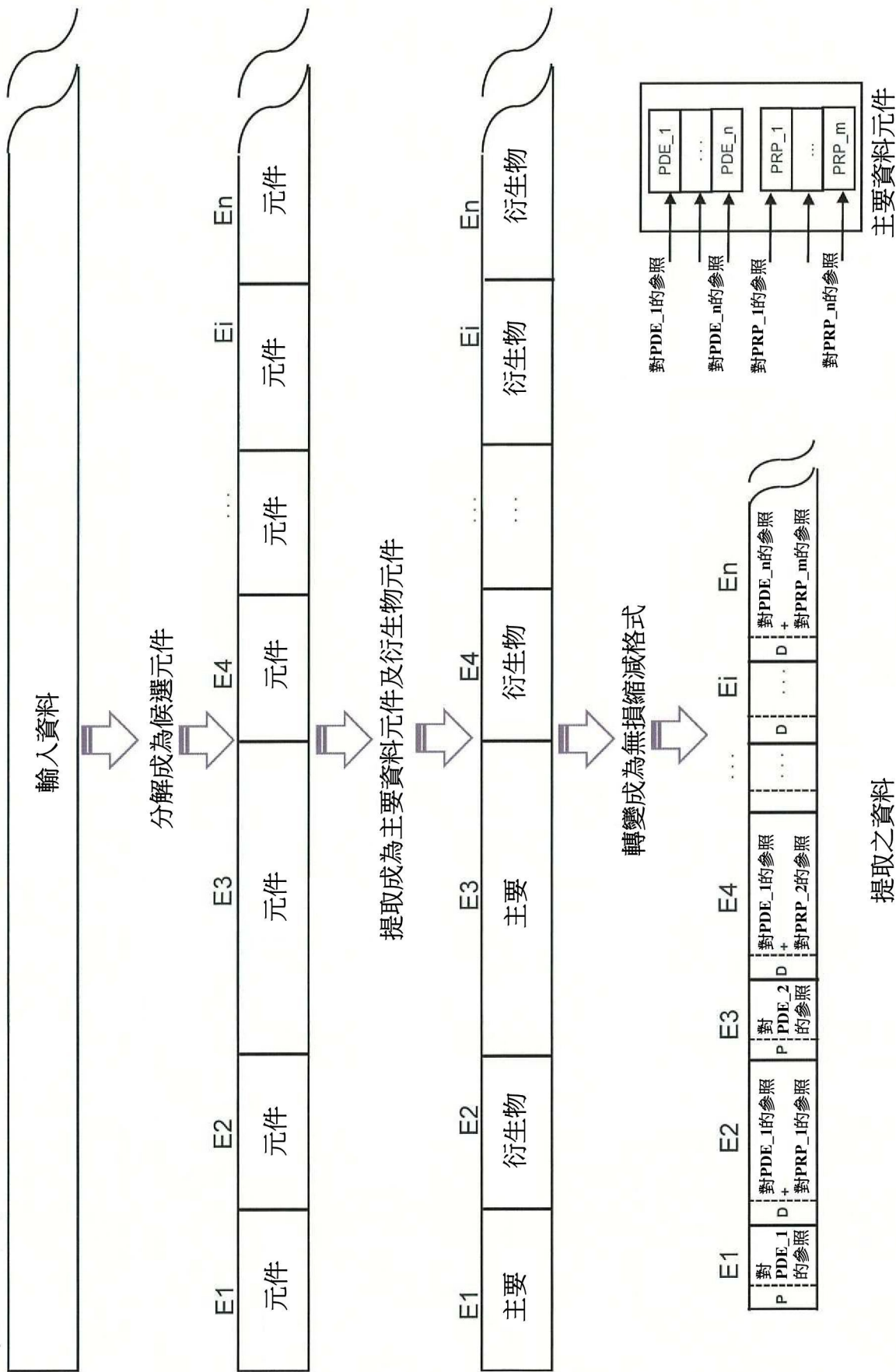
第1J 圖



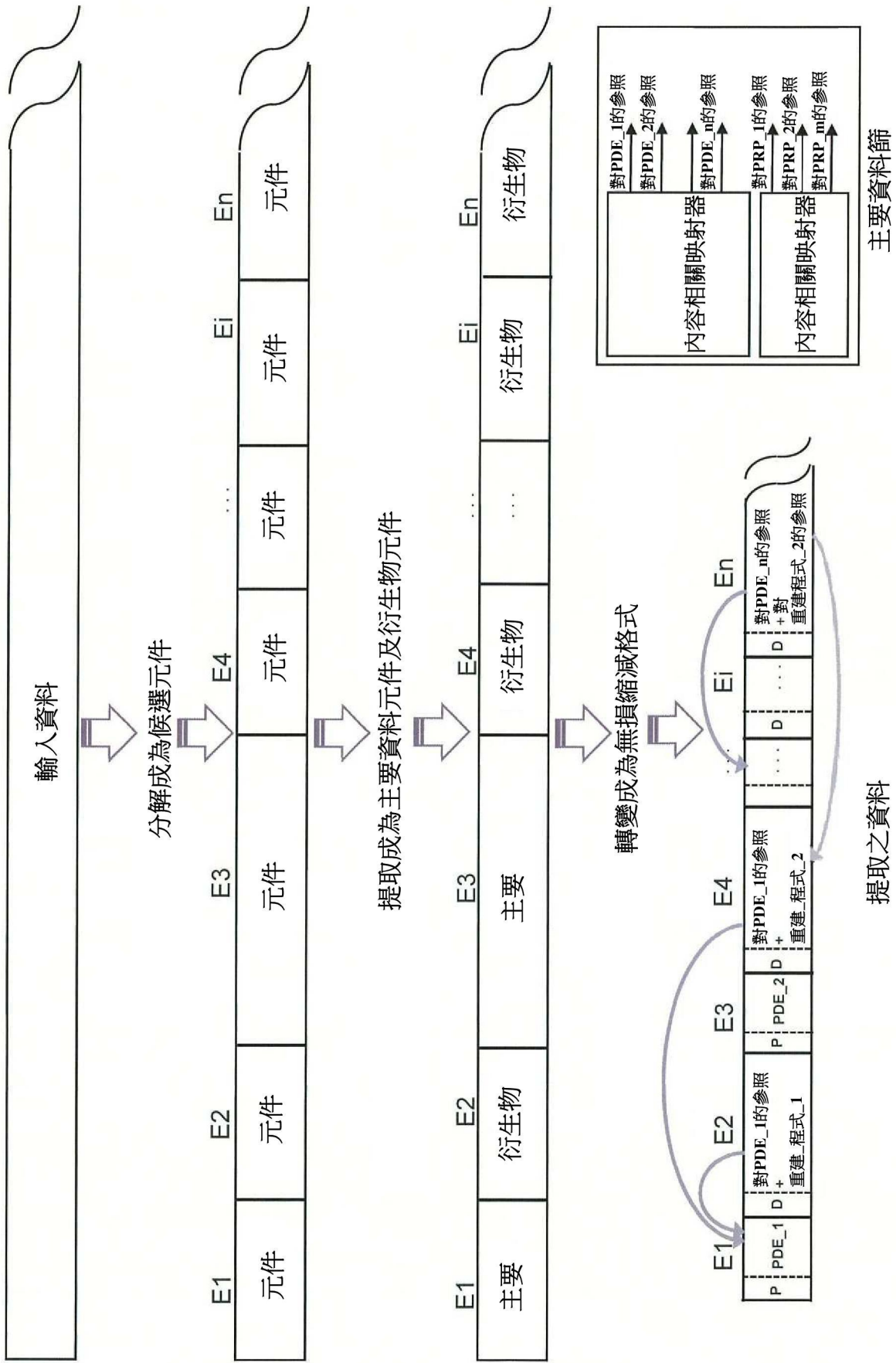
第 1K 圖



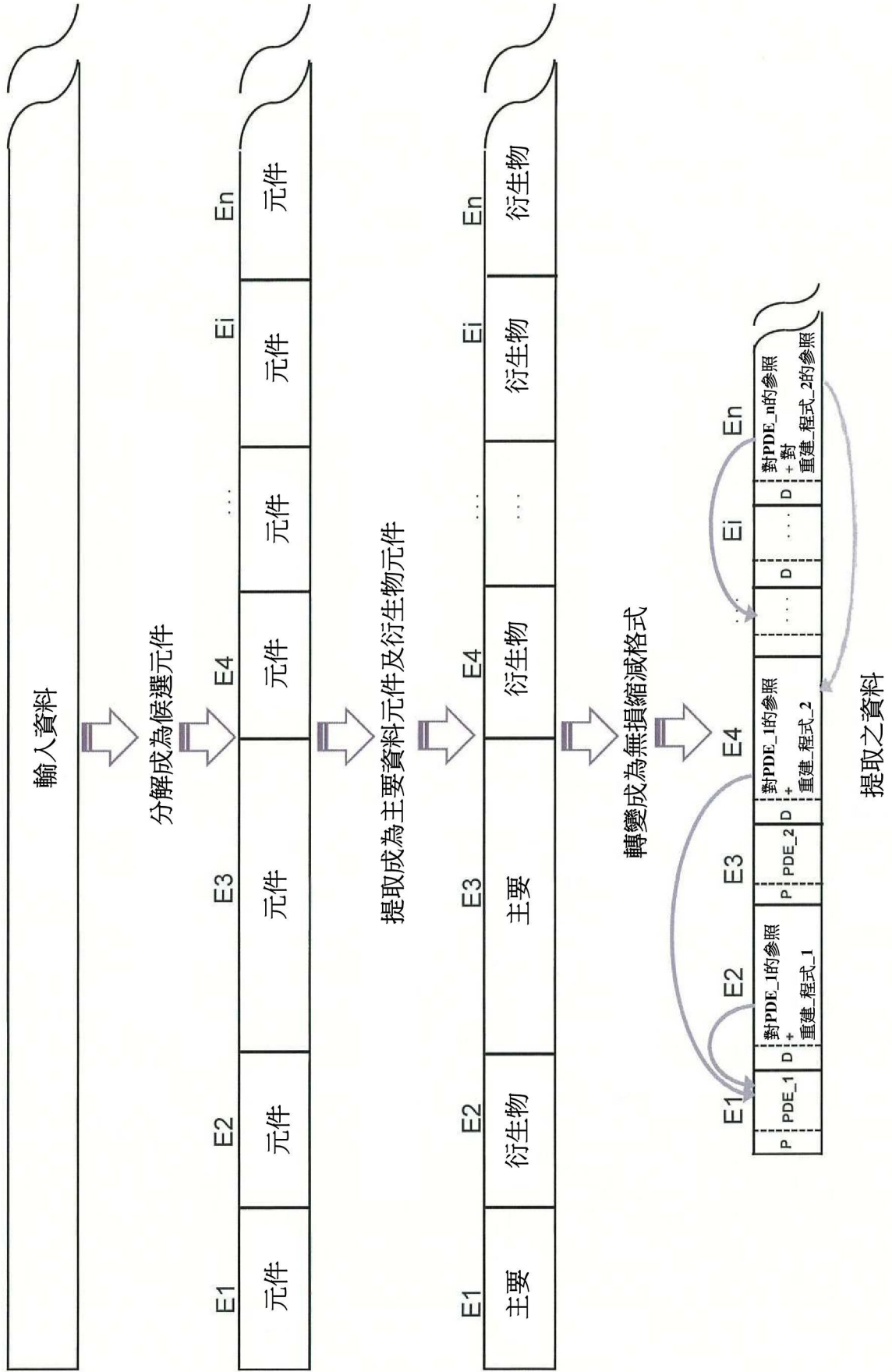
第 1L 圖



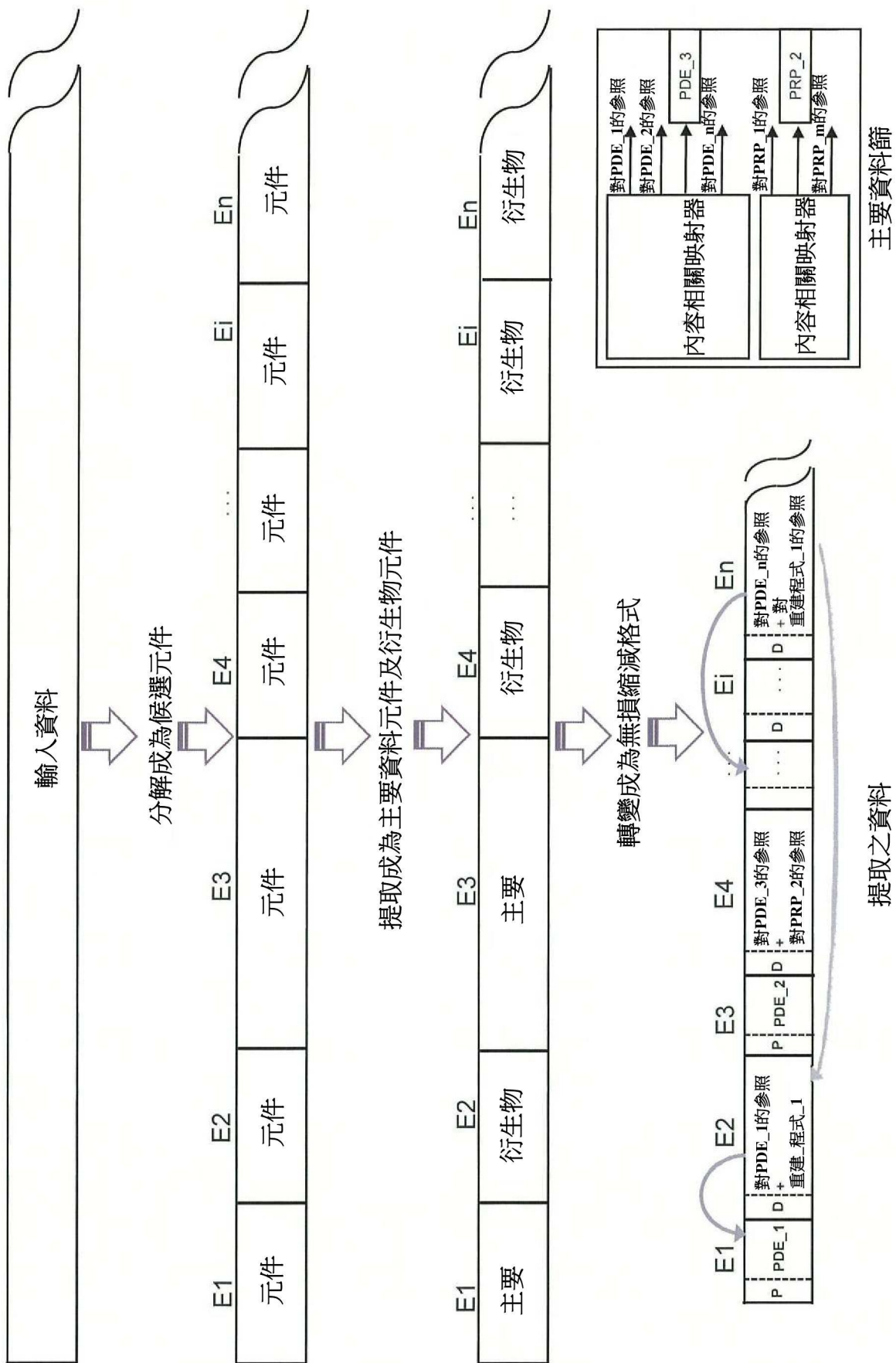
第 1M 圖



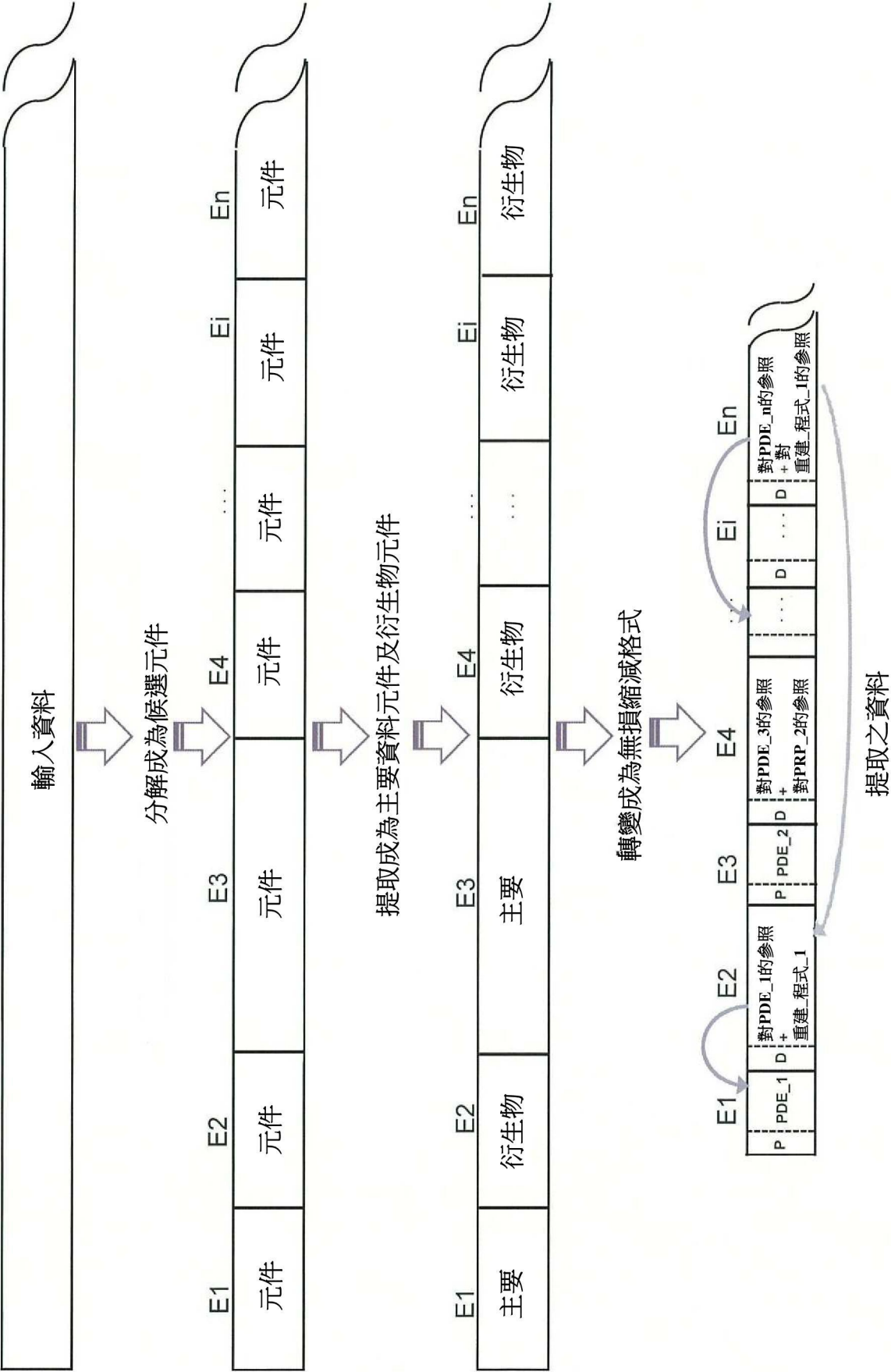
第 1N 圖



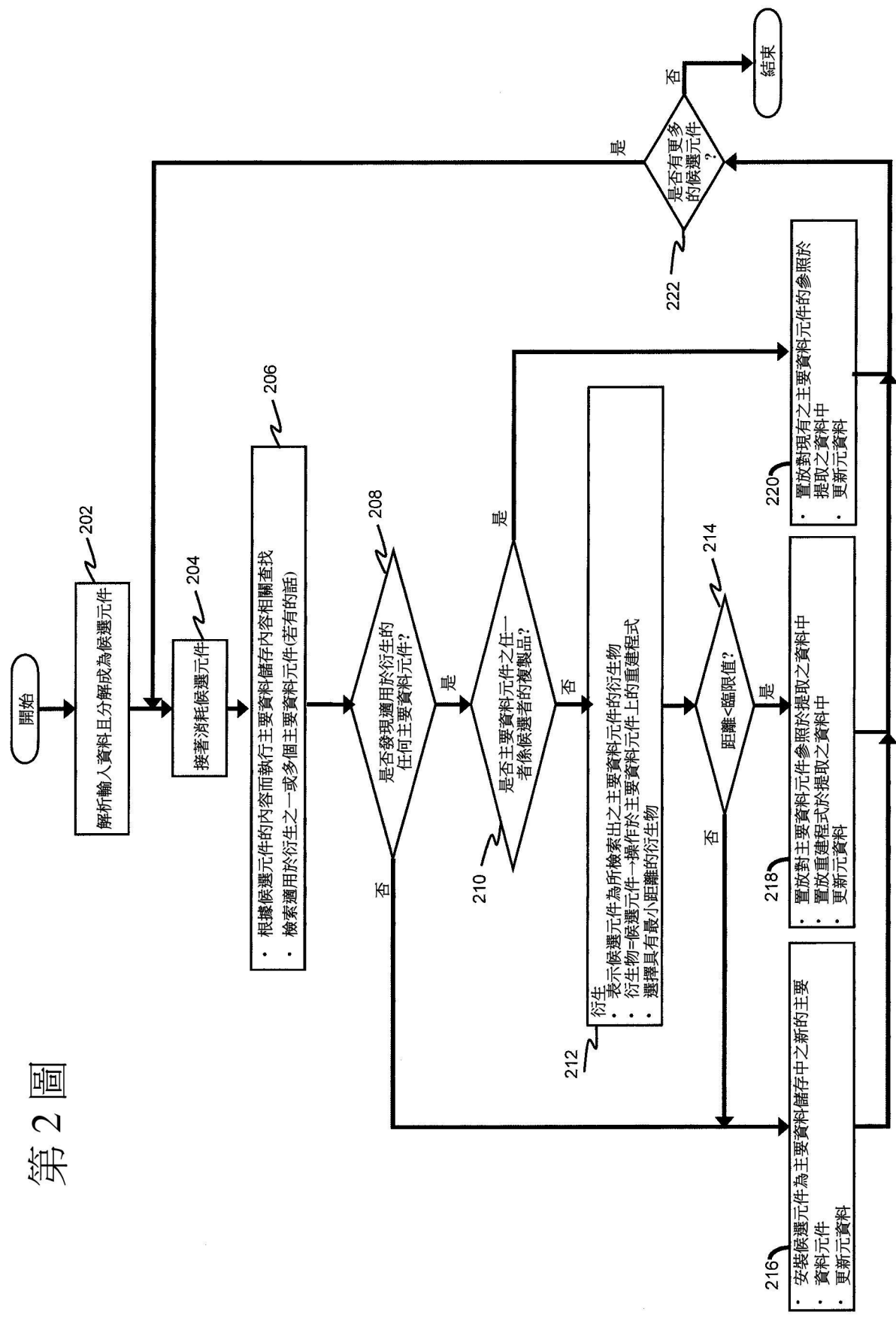
第 10 圖



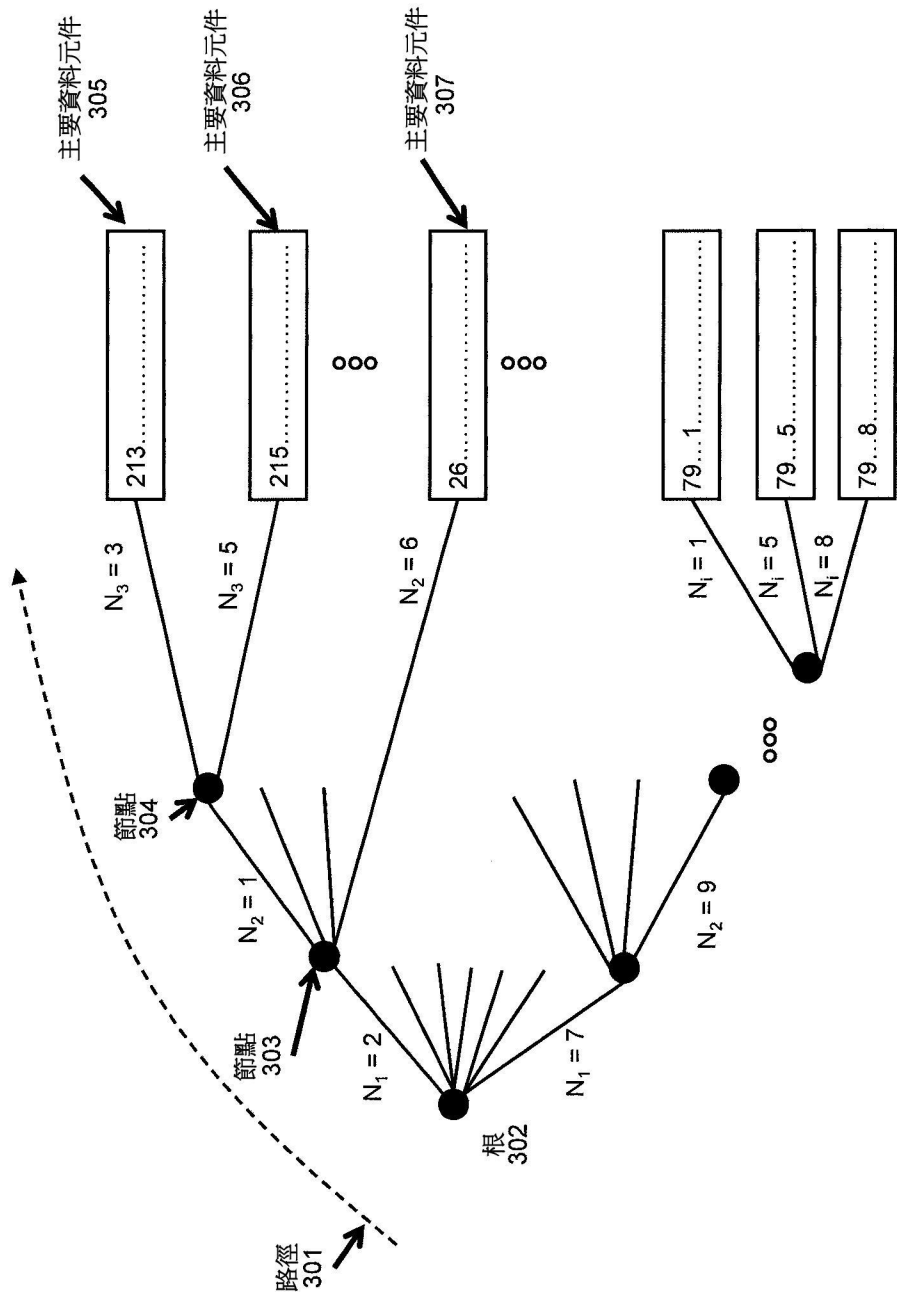
第 1P 圖



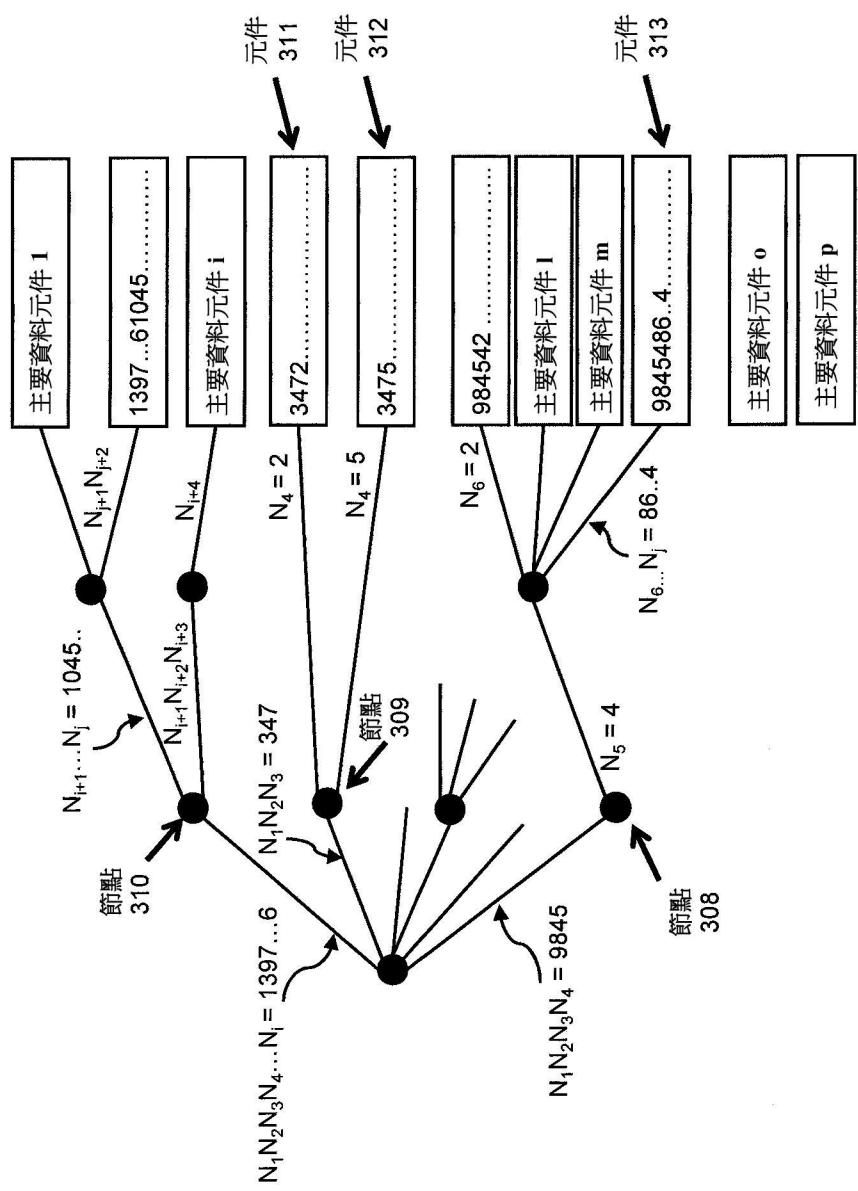
第2圖

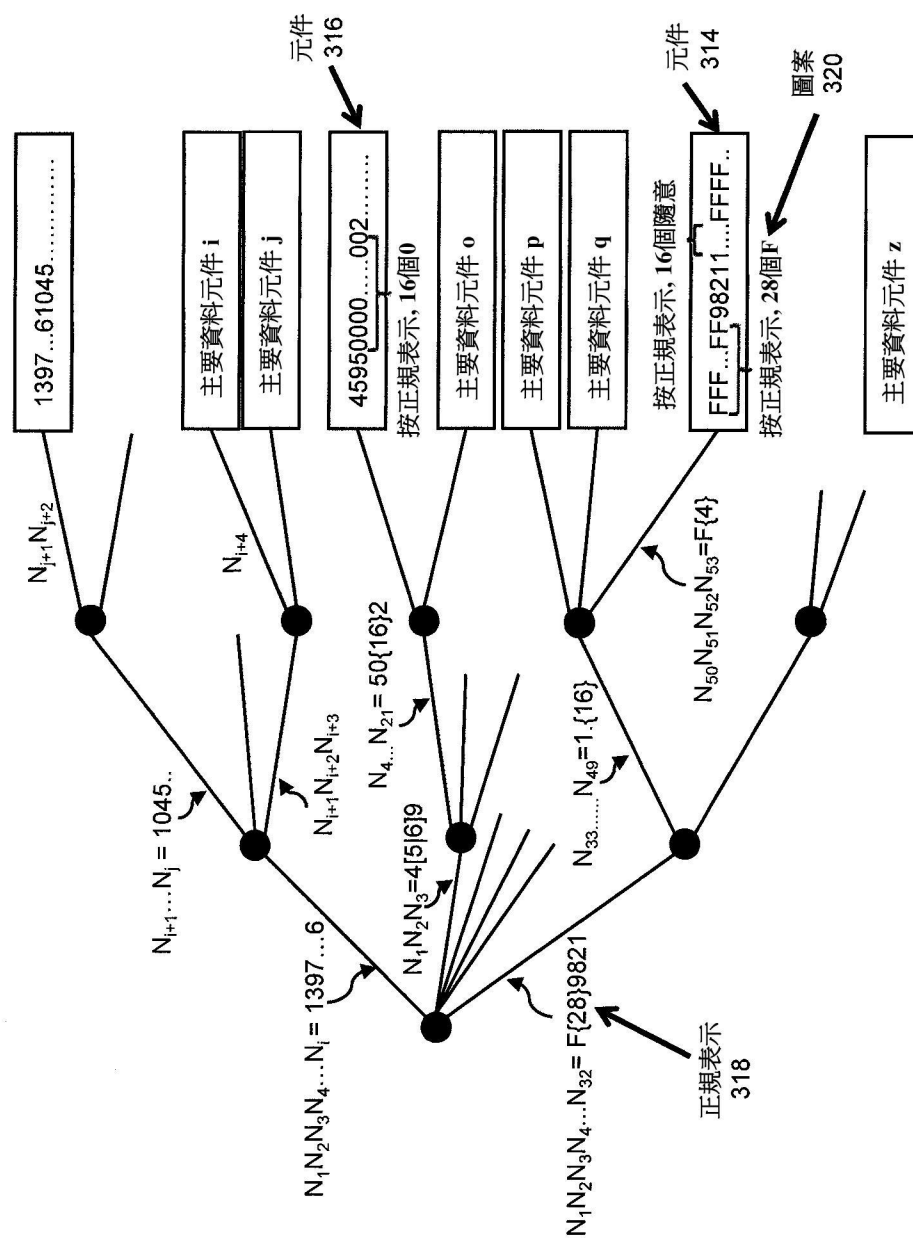


第3A圖

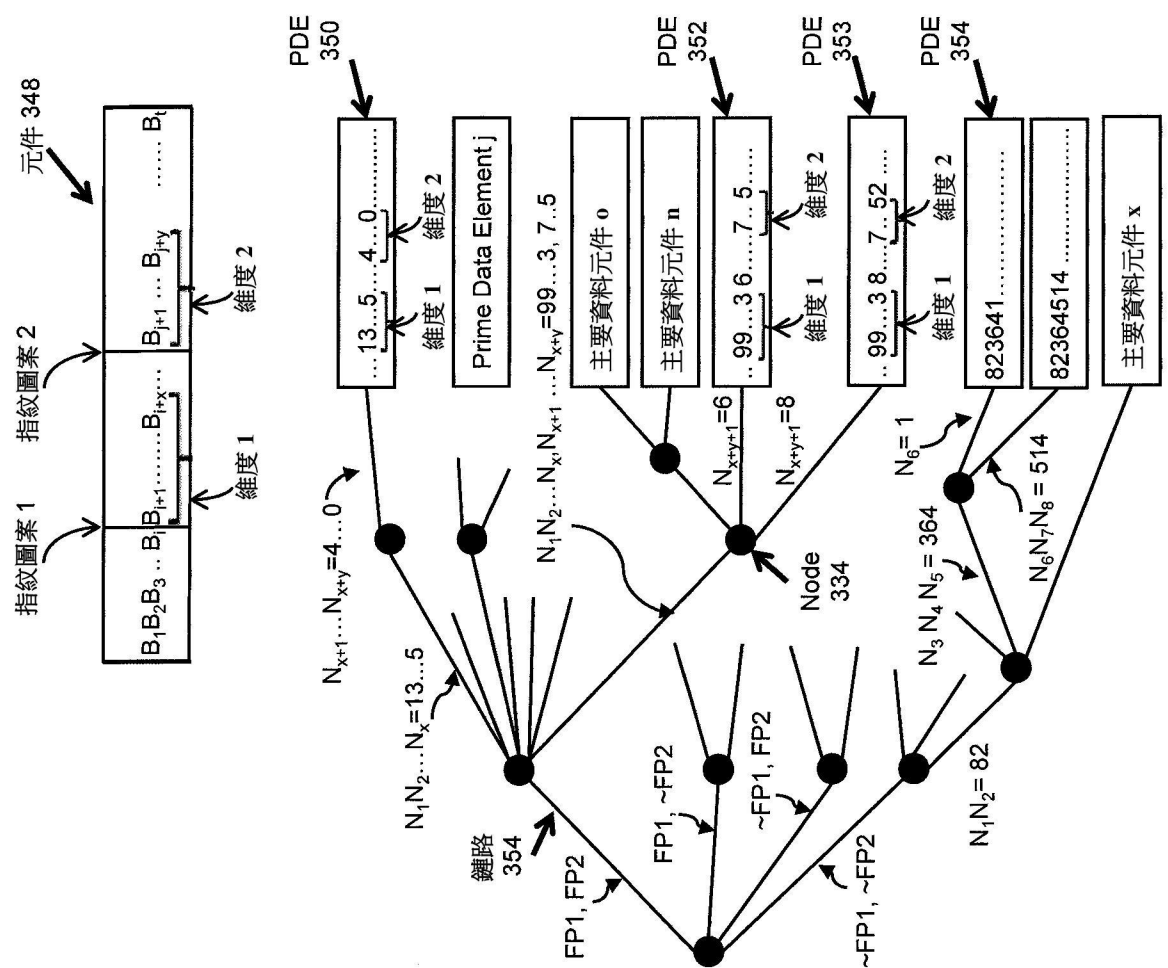


第 3B 圖





第3E圖



第 3F 圖

路徑資料			
子的數目			
N			
子 ID	不同位元組的數目	不同位元組的值	對子節點的參照資料
1	6	abef12d6743a	節點 27
2	2	dcfa	節點 16
...	...	...	...
N	4	f3231929	節點 4198

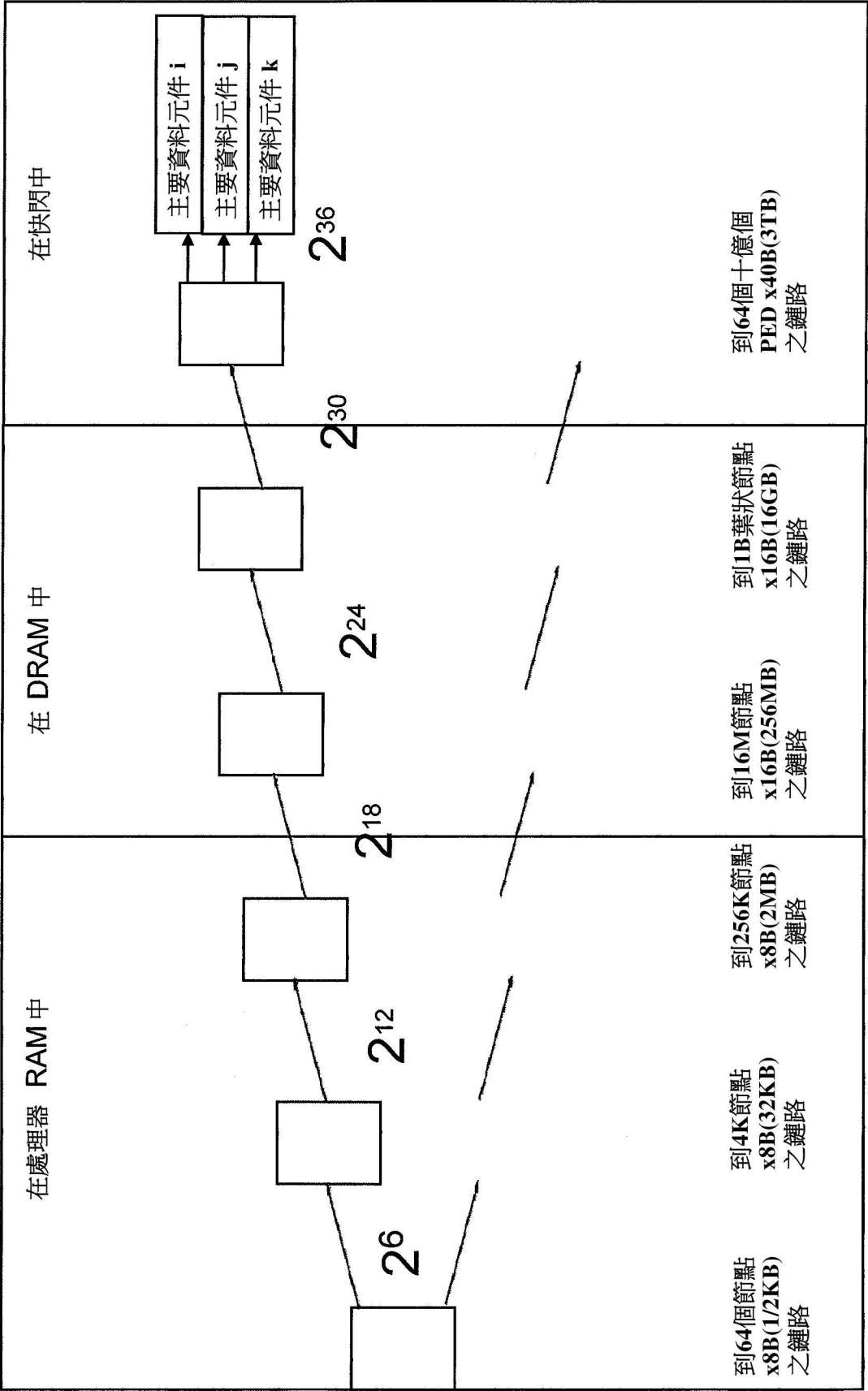
第3G圖

路徑資訊							
子的數目							
N							
子 ID	不同位元組的數目	不同位元組的值	對主要資料 元件的參照	複製品及 衍生物 的計數	用於主要資料元件 的其他元資料		
1	1	17	pde 76	4	<...>		
2	1	32	pde 4718	7	<...>		
...	...	...	...	...	...		
N	4	654aed21	pde 786	12	<...>		

第3H圖

路徑資訊								
子的數目								
N								
子 ID	不同位元組的數目	不同位元組的值	對主要資料的元件參照	導航先行位元組	複製品及衍生物的計數	用於主要資料元件的其他元資料		
1	1	17	pde 76	13476231	4	<...>		
2	1	32	pde 4718	00337650	7	<...>		
...	...	...	...	...	...	...		
N	4	654aed21	pde 786	ed189721	12	<...>		

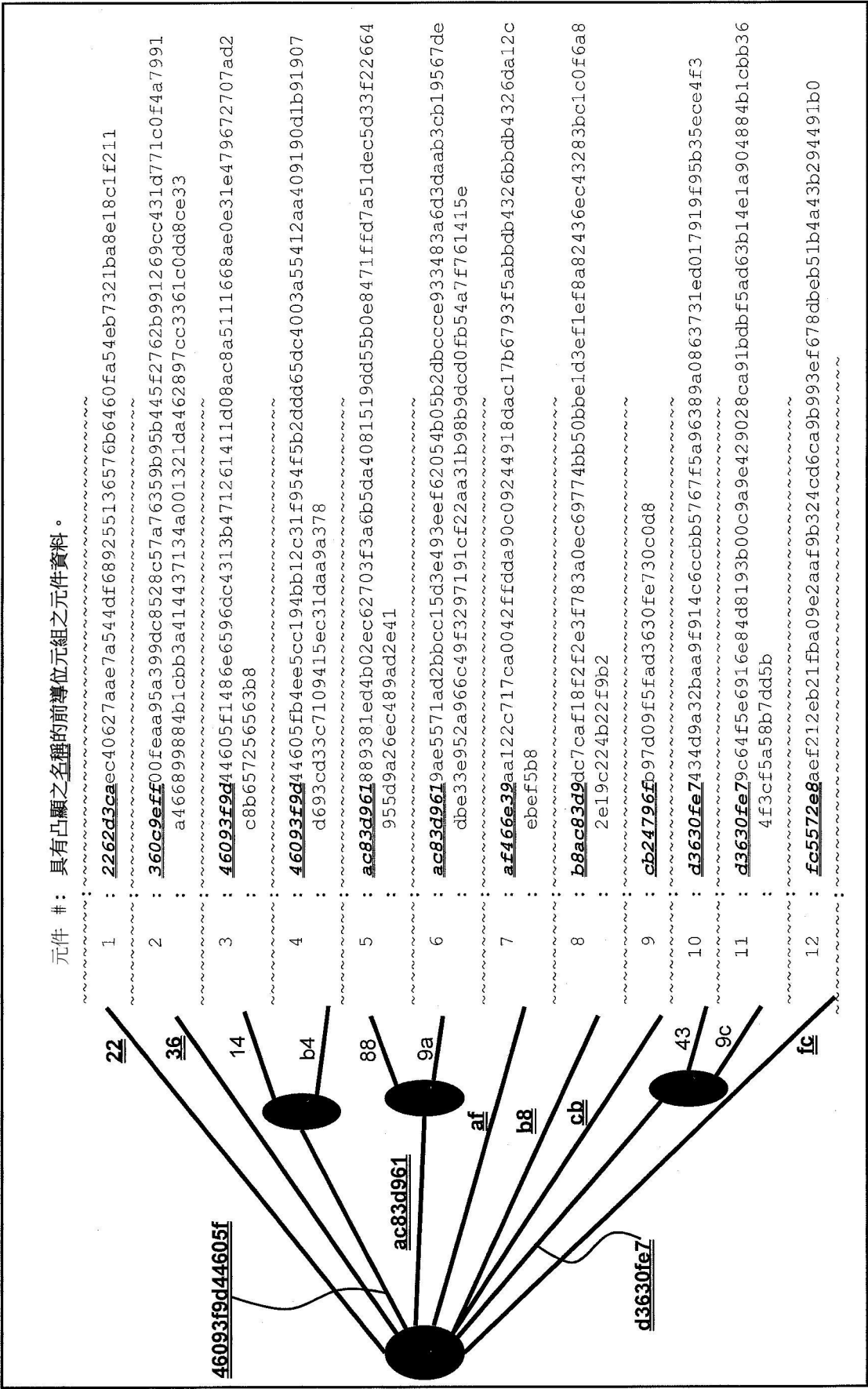
第 4 圖



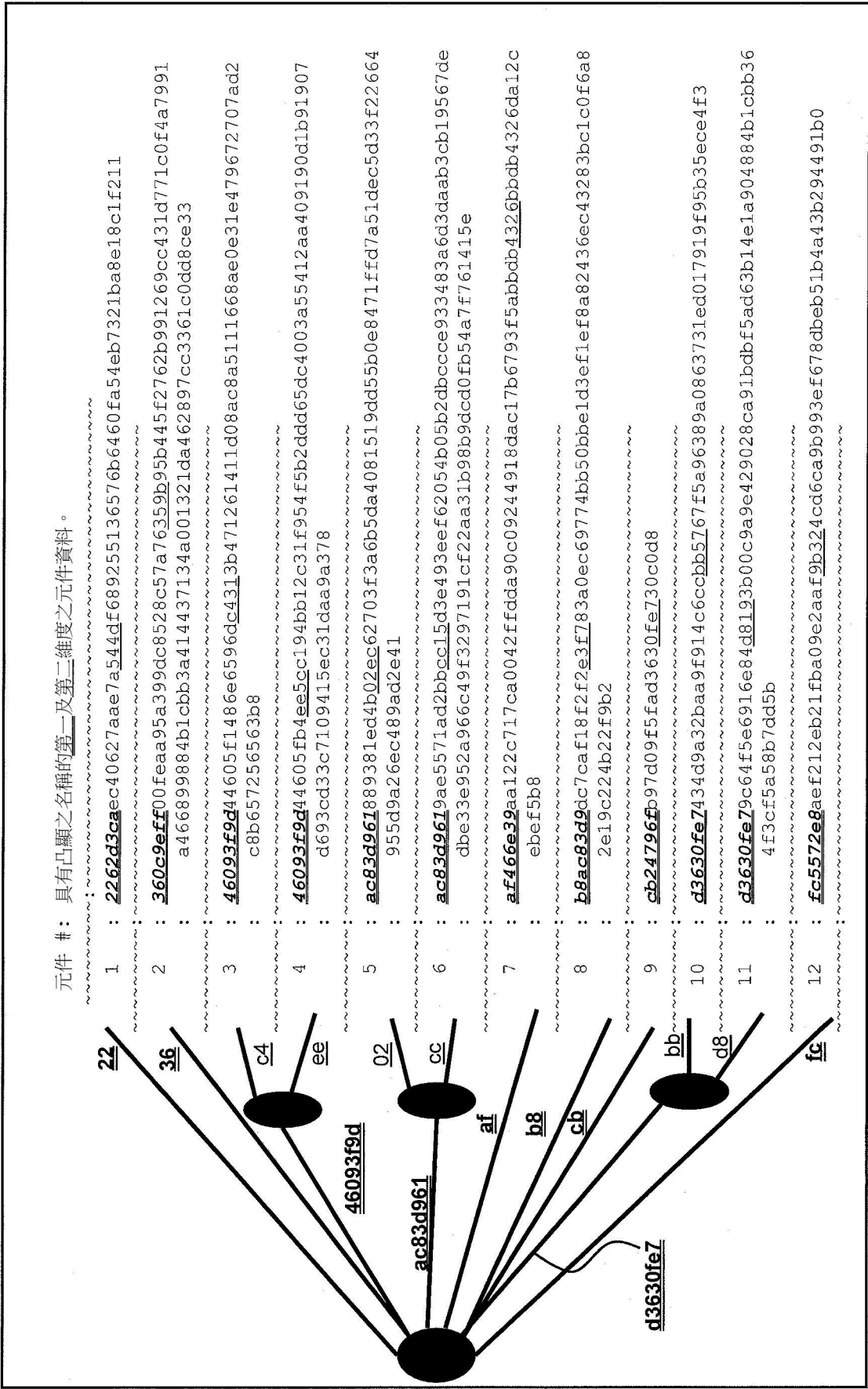
第 5A 圖

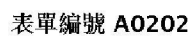
b8ac83d9dc7caf18f2f2e3f783a0ec69774bb50bbe1d3ef1ef8a82436ec43283
bc1c0f6a82e19c224b22f9b2ac83d9619ae5571ad2bbcc15d3e493eef62054b0
5b2dbccce933483a6d3daab3cb19567dedbe33e952a966c49f3297191cf22aa3
1b98b9dcd0fb54a7f761415efc5572e8aef212eb21fba09e2aaf9b324cd6ca9b
993ef678dbeb51b4a43b294491b046093f9d44605fb4ee5cc194bb12c31f954f
5b2ddd65dc4003a55412aa409190d1b91907d693cd33c7109415ec31daa9a378
2262d3caec40627aae7a544df689255136576b6460fa54eb7321ba8e18c1f211
360c9eff00feaa95a399dc8528c57a76359b95b445f2762b991269cc431d771c
0f4a7991a466899884b1cbb3a414437134a001321da462897cc3361c0dd8ce33
d3630fe7434d9a32baa9f914c6ccb5767f5a96389a0863731ed017919f95b35
ece4f346093f9d44605f1486e6596dc4313b471261411d08ac8a5111668ae0e3
1e479672707ad2c8b657256563b8af466e39aa122c717ca0042ffdda90c09244
918dac17b6793f5abbdb4326b6db4326da12cebef5b8ac83d961889381ed4b02
ec62703f3a6b5da4081519dd55b0e8471ffd7a51dec5d33f22664955d9a26ec4
89ad2e41d3630fe79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad63b1
4e1a904884b1cbb364f3cf5a58b7dd5bcb24796fb97d09f5fad3630fe730c0d8

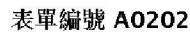
第5B圖



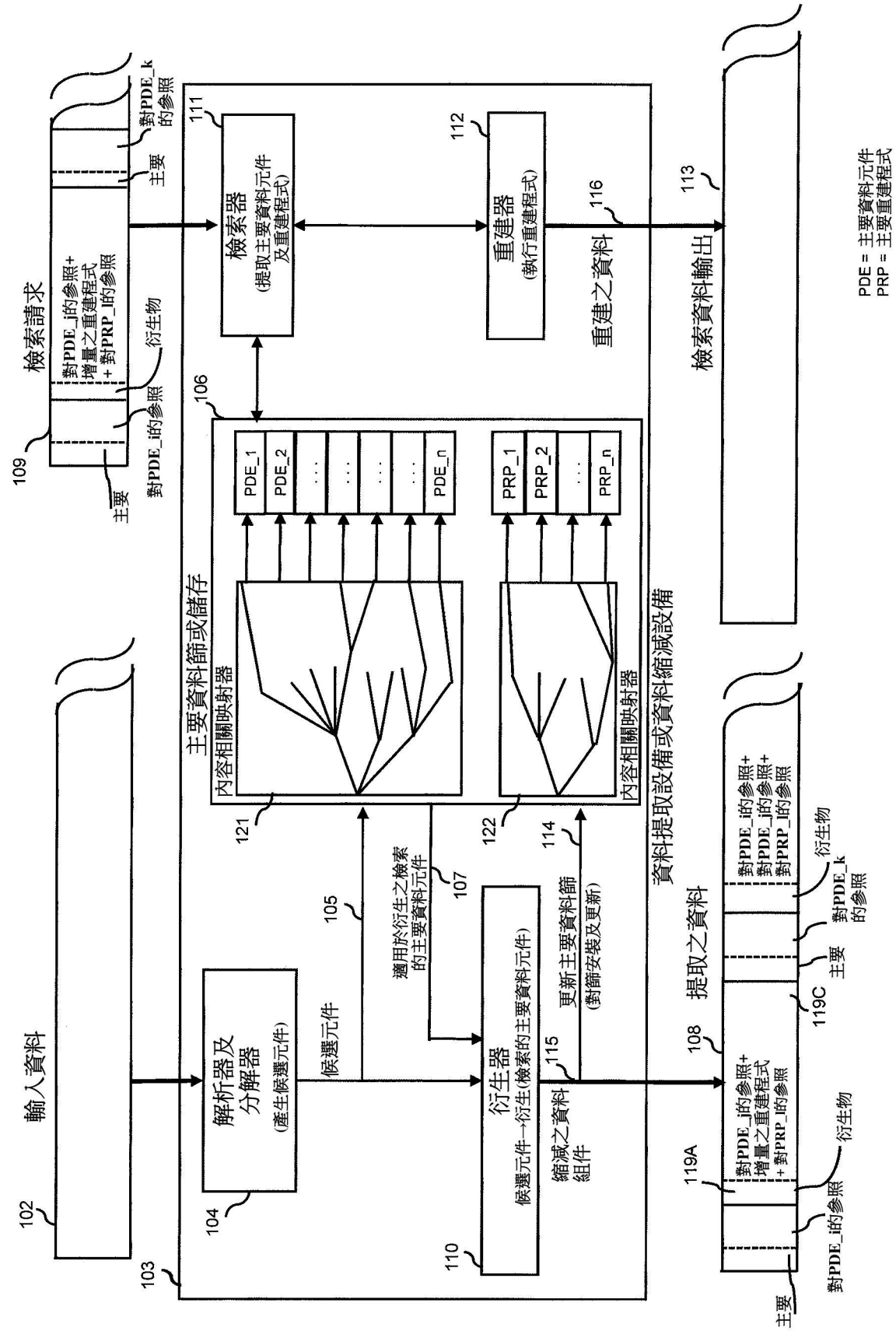
第5C圖







第6C圖



第7A圖 用於由重建程式所使用的運算之指令編碼的範例

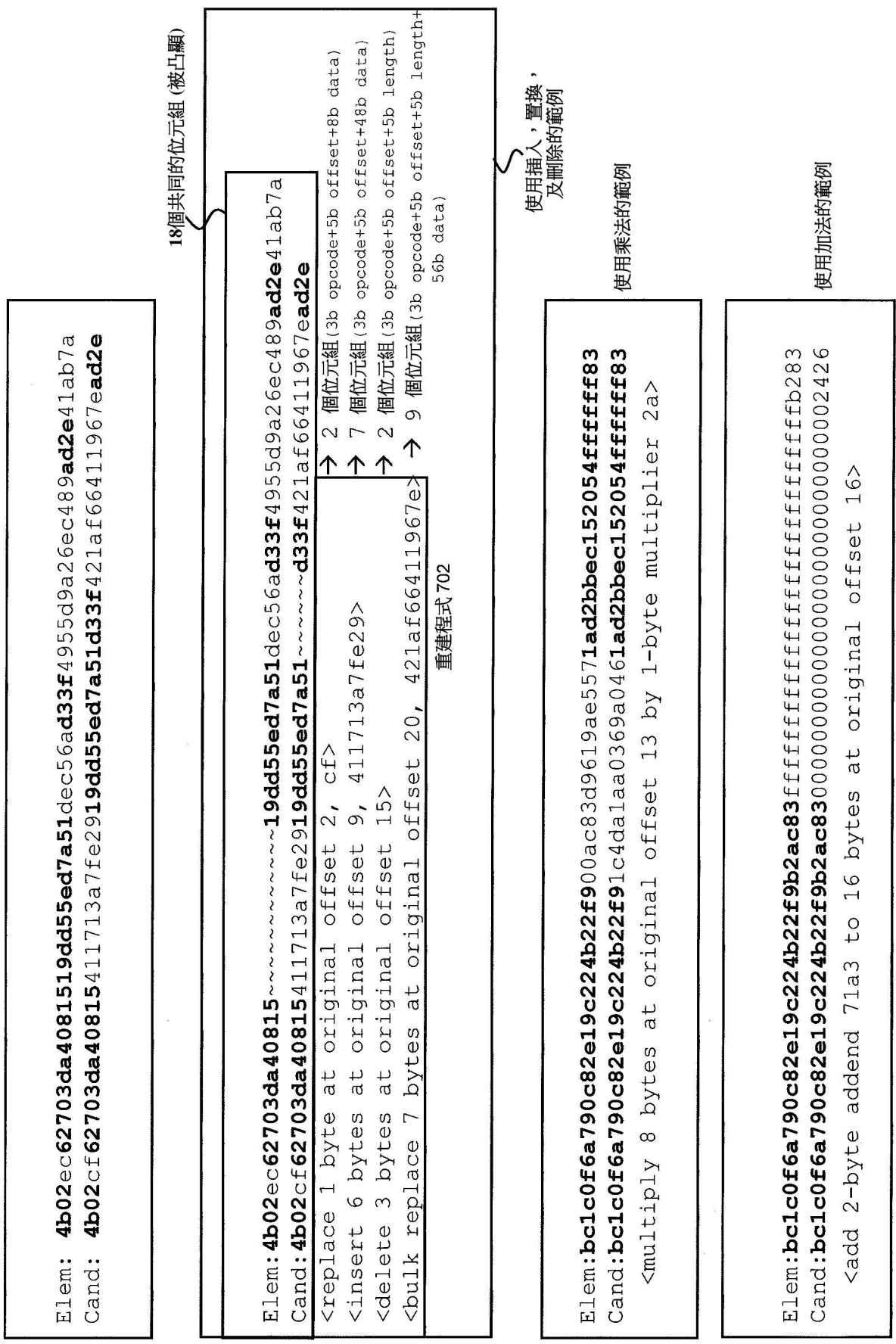
運算	運算碼	運算元
插入	0	原始偏移，插入長度，插入資料 ('長度'位元組)
刪除	1	原始偏移，長度
置換	2	原始偏移，置換資料 (1位元組)
大批置換	3	原始偏移，長度，置換資料 ('長度'位元組)
附加	4	長度，附加資料 ('長度'位元組)
相乘	5	原始偏移，乘積長度，乘數長度，乘數值
相加	6	原始偏移，和長度，加數長度，加數值
擴展運算碼	7	參閱擴展運算表

擴展運算	運算碼	範圍	運算及運算元
算術	0	單一PDE	運算，原始偏移，運算長度，運算者長度，運算者
多個元件置換 (從第二元件以資料置換)	1	多個PDE	第二元件處置，第二元件偏移，原始偏移，置換長度
多個元件插入(自第二元件插入資料)	2	多個PDE	第二元件處置，第二元件偏移，原始偏移，插入長度
直列呼叫及執行主要重建程式 (自指明之位址提取重建程式且 在單一主要資料元件之上執行)	3	單一PDE	對所呼叫之主要重建程式的參照
提取，修正，及執主要重建程式 (自指明之位址提取重建程式，施 加運算於隨後之窗口以取得新的 重建程式，且然後，執行它)	4	重建程式	對主要重建程式的參照，窗口
保留	5-7	保留	保留

欄類型	欄尺寸
運算碼或運算	3 位元
偏移	5 位元
長度	5 位元
參照	24 位元
處置	2 位元

注意：PDE=主要資料元件

圖 7B 第



第 8A 圖

bc1c0f6a82e19c224b22f9b2ac83d9619ae5571ad2bbcc15d3e493eef62054b0	: 資料塊 1
fc29344c742bb3e45c197e655912553bd07ab2a91e9d3d2a3feddd3ed2d442b6	: 資料塊 2
993ef678dbeb51b4a43b294491b046093f9d44605fb4ee5cc194bb12c31f954f	: 資料塊 3
65bc3565a3d9457b656eb3b4a72da8302ec5fb8dc7eeb77c577de08d2b46d3f4	: 資料塊 4
2262d3caec40627aae7a544df689255136576b6460fa54eb7321ba8e18c1f211	: 資料塊 5
0f4a7991a466899884b1d3630fe79c64f5e601321da462897cc3361c0dd8ce33	: 資料塊 6
1e479672707ad2c8b657256563b8af466e39aa122c717ca0042ffdda90c09244	: 資料塊 7
918dac17b6793f5abbdb4326bbdb4326da12cebef5b8ac83d961889381ed4b02	: 資料塊 8
ec62703f3a6b5da4081519dd55b0e8471ffd7a51dec5d33f22664955d9a26ec4	: 資料塊 9
89ad2e3ad3630fe79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad63b1	: 資料塊 10
~~~~~ 42個百萬個資料塊在中間 ~~~~~	
ec62703f3a6b5da4081519dd55b0e8471ffd7a51dec5d33f22664955d9a26ec4	: 資料塊 42,000,011
bc1c0f6a82790ce19c224b22f9b2ac83d9619ae5571ad2bbec152054b0aeb712	: 資料塊 42,000,012
67b289ad2e41d3630fe79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad	: 資料塊 42,000,013
0f4a7991a466899884b1d3630fe79c64f5e6916e84d8193b00c9a9e429028ca9	: 資料塊 42,000,014



第 8C 圖

元件	資料塊	名稱的 MSB	具有凸顯之名稱的 <b>第一、第二及第三</b> 維度之元件
1	2	<b>12553b1e9d197e</b> : fc29344c742bb3e45c197e6559 <b>12553bd07ab2a91e9d3d2a3feddd3ed2d442b6</b>	
2	3	<b>46093f44460b4a4</b> : 993ef678dbeb51b4a43b294491b0 <b>46093f9d44605fb4ee5cc194bb12c31f954f</b>	
		Numerous elements installed	
13,143	5	<b>544df636577aae</b> : 2262d3caec40627aae7a <b>544df689255136576b6460fa54eb7321ba8e18c1f211</b>	
13,144	4	<b>6eb3b42ec565a3</b> : 65bc3565a3d9457b65 <b>6eb3b4a72da8302ec5fb8dc7eeb77c577de08d2b46d3f4</b>	
		Numerous elements installed	
24,789	9	<b>703f3a19dd6b5d</b> : ec62 <b>703f3a6b5da4081519dd55b0e8471ffd7a51dec5d33f22664955d9a26ec4</b>	
		Numerous elements installed	
187,125	8	<b>ac83d988934326</b> : 918dac17b6793f5abbdb4326bbdb4326da12cebef5b8 <b>ac83d961889381ed4b02</b>	
187,126	1	<b>ac83d9cc1582e1</b> : bc1c0f6a82e19c224b22f9b2 <b>ac83d9619ae571ad2bbcc15d3e493eef62054b0</b>	
		Numerous elements installed	
1,247,565	7	<b>af466e717c7ad2</b> : 1e479672707ad2c8b657256563b8 <b>af466e39aa122c717ca0042ffdda90c09244</b>	
		Numerous elements installed	
9,299,997	6	<b>d3630fa4628998</b> : 0f4a7991a466899884b1 <b>d3630fe79c64f5e601321da462897cc3361c0dd8ce33</b>	
9,299,998	10	<b>d3630fe69189ad</b> : 89ad2e3ad <b>d3630fe79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad63b1</b>	
16,000,010		element 16,000,010	



第 8D 圖

資料塊 42,000,011:

ec62703f3a6b5da4081519dd55b0e8471ffd7a51dec5d33f22664955d9a26ec4

資料塊 42,000,012:

bc1c0f6a82790ce19c224b22f9b2ac83d9619ae5571ad2bbec152054b0aeb712

衍生對元件 187,125:

918dac17b6793f5abbbdb4326b6bdb4326da12cebef5b8ac83d961889381ed4b02~~~~~  
bc1c0f6a790c82e19c224b22f9b2~~~~~ac83d9619ae5571ad2bbec152054b0aeb712  
    <replace 14 bytes at original offset 0, bc1c0f6a790c82e19c224b22f9b2>  
    <delete 8 bytes at original offset14>  
    <replace 6 bytes at original offset 18,9ae5571ad2bb>  
    <insert 8 bytes at original offset 24, ec152054b0aeb712>

衍生對元件 187,126:

bc1c0f6a82~~~~e19c224b22f9b2ac83d9619ae5571ad2bbec15d3e493eef62054b0  
bc1c0f6a82790ce19c224b22f9b2ac83d9619ae5571ad2bbec15~~~~~2054b0aeb712  
    <insert 2 bytes at original offset 4, 790c>  
    <replace 1 byte at original offset 22, ec>  
    <delete 5 bytes at original offset 24>  
    <append 3 bytes at end of accumulated bytes, aeb712>



第 8E 圖

資料塊 42,000,013:

67b289ad2e3a**d3630fe**79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad

衍生對元件 9,299,998:

~~~~89ad2e3ad**d3630fe**79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad63b1  
67b289ad2e3a**d3630fe**79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad
<insert 2 bytes at original offset 0, 67b2>

資料塊 42,000,014:

0f4a7991a466899884b1**d3630fe**79c64f5e6916e84d8193b00c9a9e429028ca9

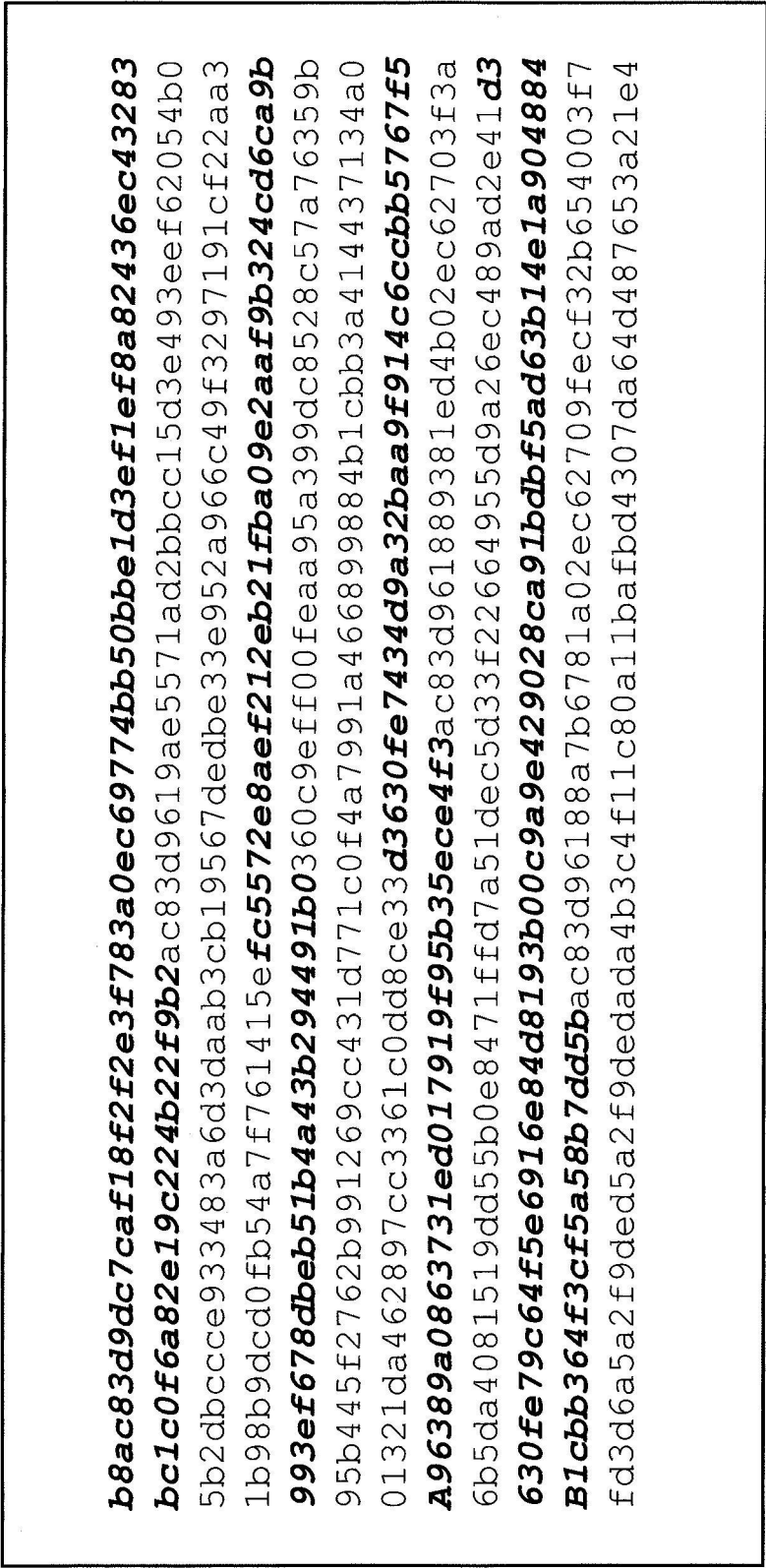
衍生對元件 9,299,997 and 9,299,998:

0f4a7991a466899884b1**d3630fe**79c64f5e601321da462897cc3361c0dd8ce33
89ad2e3a**d3630fe**79c64f5e6916e84d8193b00c9a9e429028ca91bdbf5ad63b1

0f4a7991a466899884b1**d3630fe**79c64f5e6916e84d8193b00c9a9e429028ca9

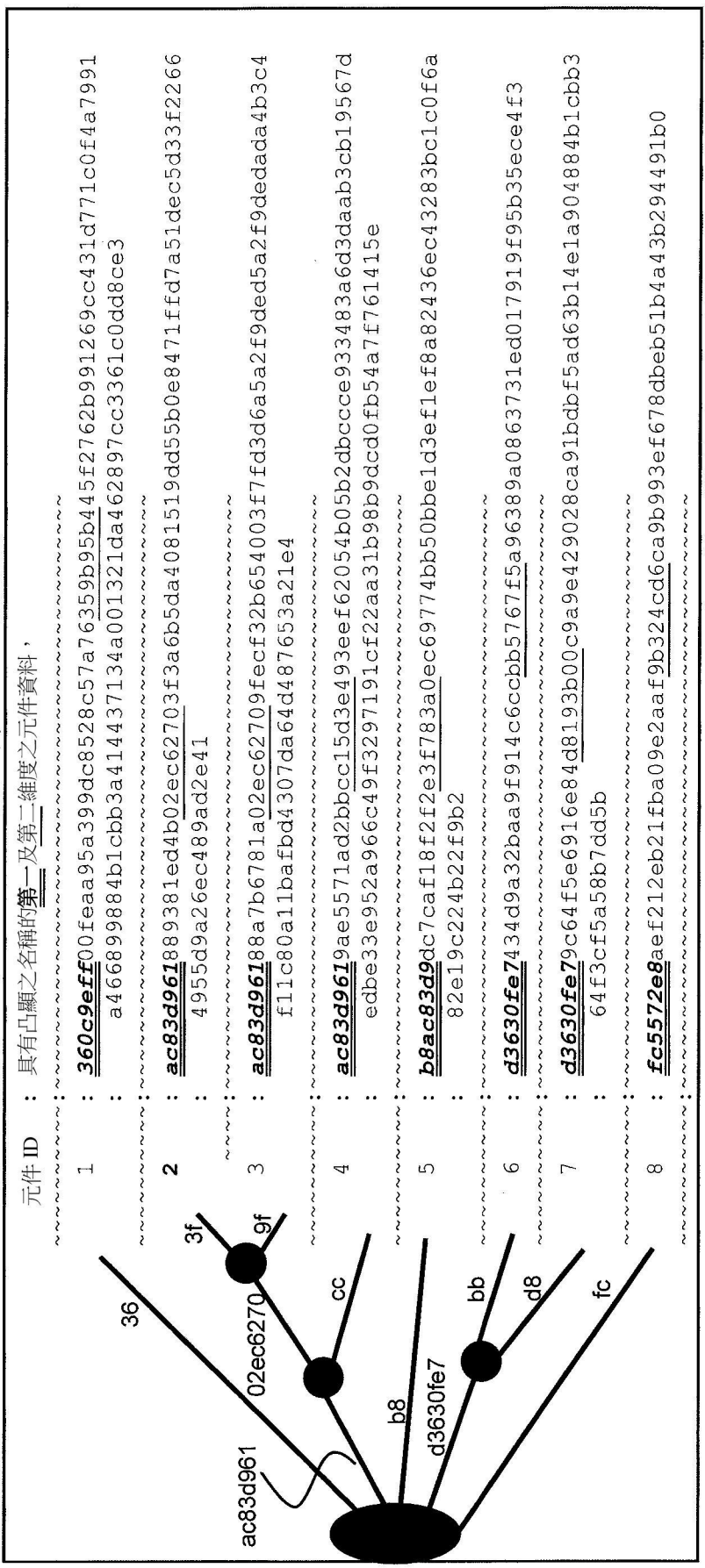
(start with element 9,299,997)
<multi-element insert, element 9,299,998, original offset 13,
2<sup>nd</sup> element offset 7, length 19>

第 9A 圖

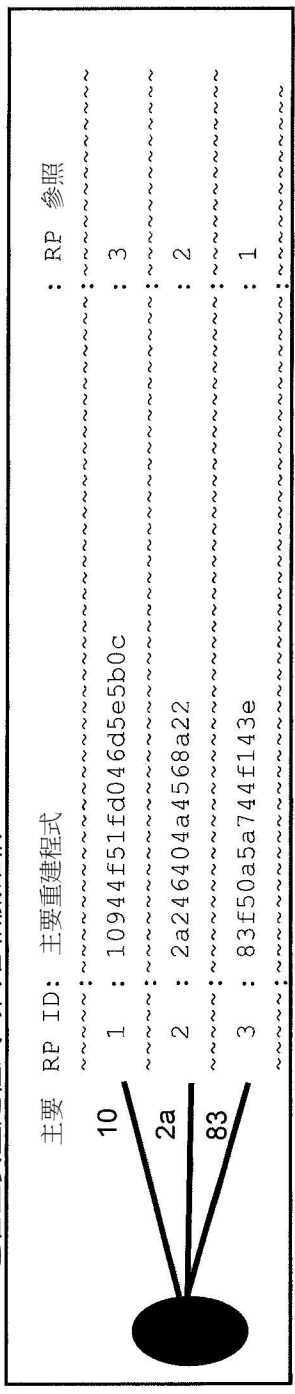


第 9B 圖

節



包含主要重建程式的內容相關結構



ac83d96188a7b6781a02ec62707aecf32b654076231a03f76a5a2f9ded5a2f9deda4b3c4f11c80a11bafbd4307da64d487653a21e4

編碼的重建程式RP

```

: 46bd14c4098aea4e3407eed4

```

: RP 参照

```

1 : 10944f51fd046d5e5b0c : 3
2 : 2a246404a4568a22 : 2
3 : 46bd14c098aa4e3407eed4 : 4
4 : 83f50a5a744f143e : 1

```

12個位元組RP

46bd14c4098aea4e3407eed4 被插入至主要重建程式篩

: RP 参照

```

1 : 10944f51fd046d5e5b0c : 3
2 : 2a246404a4568a22 : 2
3 : 46bd14c4098aaa4e3407eed4 : 4
4 : 7a0e73de80 : 6
5 : 83f50a5a744f143e : 1
6 : 90a7f84ba5bf031c6029fe12e96fc0c718 : 5

```

12個位元組RP

46bd14c4098aea4e3407eed4
在主要重建程式篩中所搜尋出

第 10A 圖

重建程式：

<insert 2 bytes at original offset 4, 790c>
<replace 1 byte at original offset 22, ec>
<delete 5 bytes at original offset 24>
<append 3 bytes at end of accumulated bytes, aeb783>

元件 187,126:

bc1c0f6a82e19c224b22f9b2ac83d9619ae5571ad2bbcc15d3e493eef62054b0

重建程式之執行：

<insert 2 bytes at original offset 4, 790c>
→ bc1c0f6a**790c**82e19c224b22f9b2ac83d9619ae5571ad2bbcc15d3e493eef62054b0

<replace 1 byte at original offset 22, ec>
→ bc1c0f6a790c82e19c224b22f9b2ac83d9619ae5571ad2bb**ec**15d3e493eef62054b0

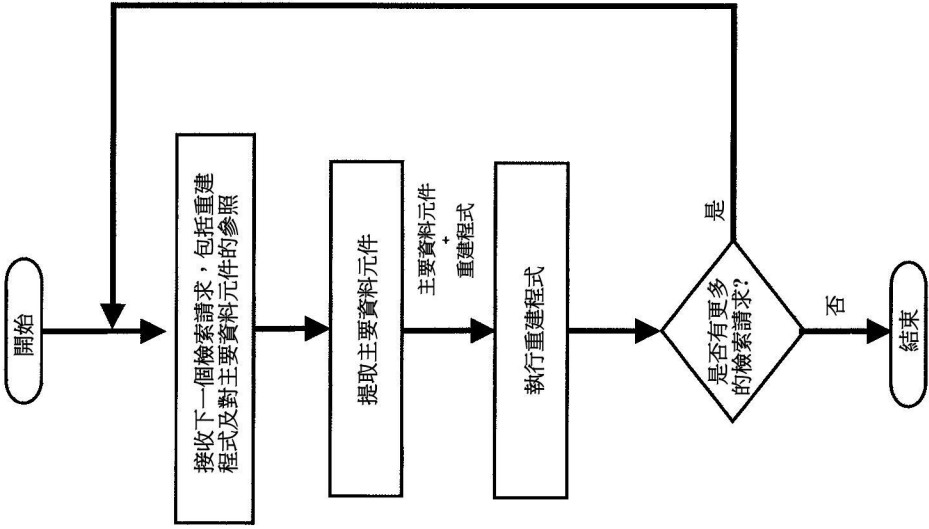
<delete 5 bytes at original offset 24>
→ bc1c0f6a790c82e19c224b22f9b2ac83d9619ae5571ad2bbec15d3e493eef62054b0

<append 3 bytes at end of accumulated bytes, aeb783>
→ bc1c0f6a790c82e19c224b22f9b2ac83d9619ae5571ad2bbec152054b0**aeb783**

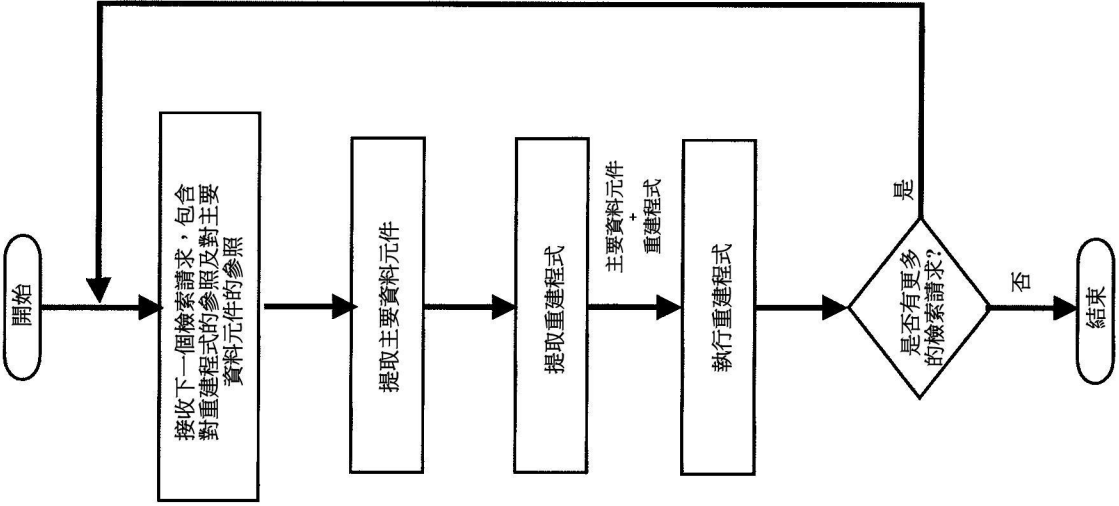
資料塊 42,000,012:

bc1c0f6a790c82e19c224b22f9b2ac83d9619ae5571ad2bbec152054b0aeb783

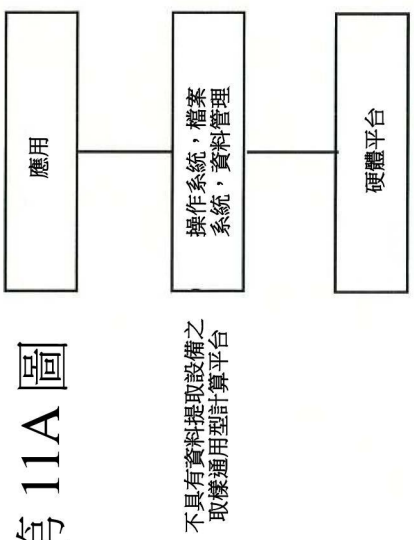
第 10B 圖



第 10C 圖

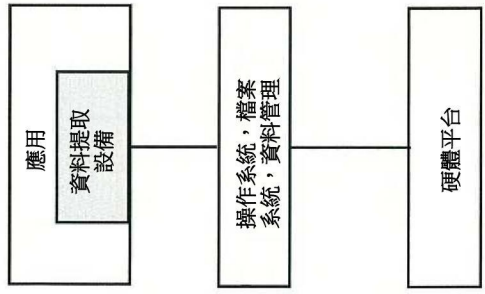


第 11A 圖

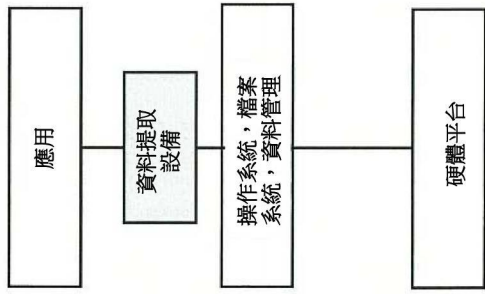


不具有資料提取設備之
取樣通用型計算平台

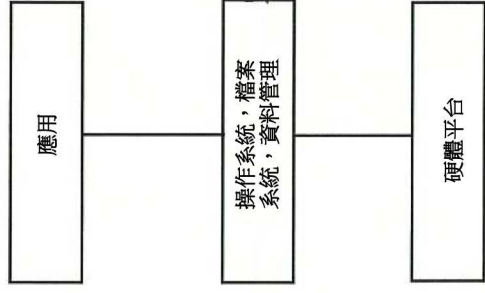
第 11B 圖



第 11C 圖

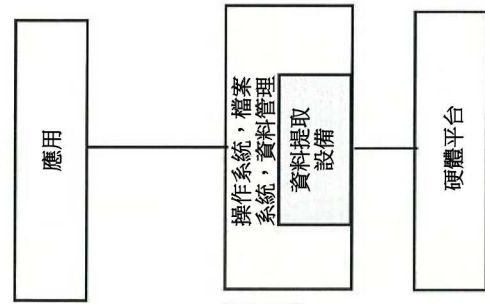


第 11D 圖

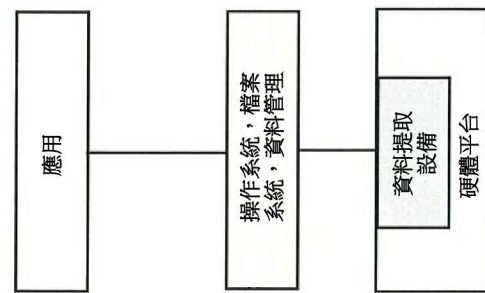


以資料提取設備強化
之取樣通用型計算平台

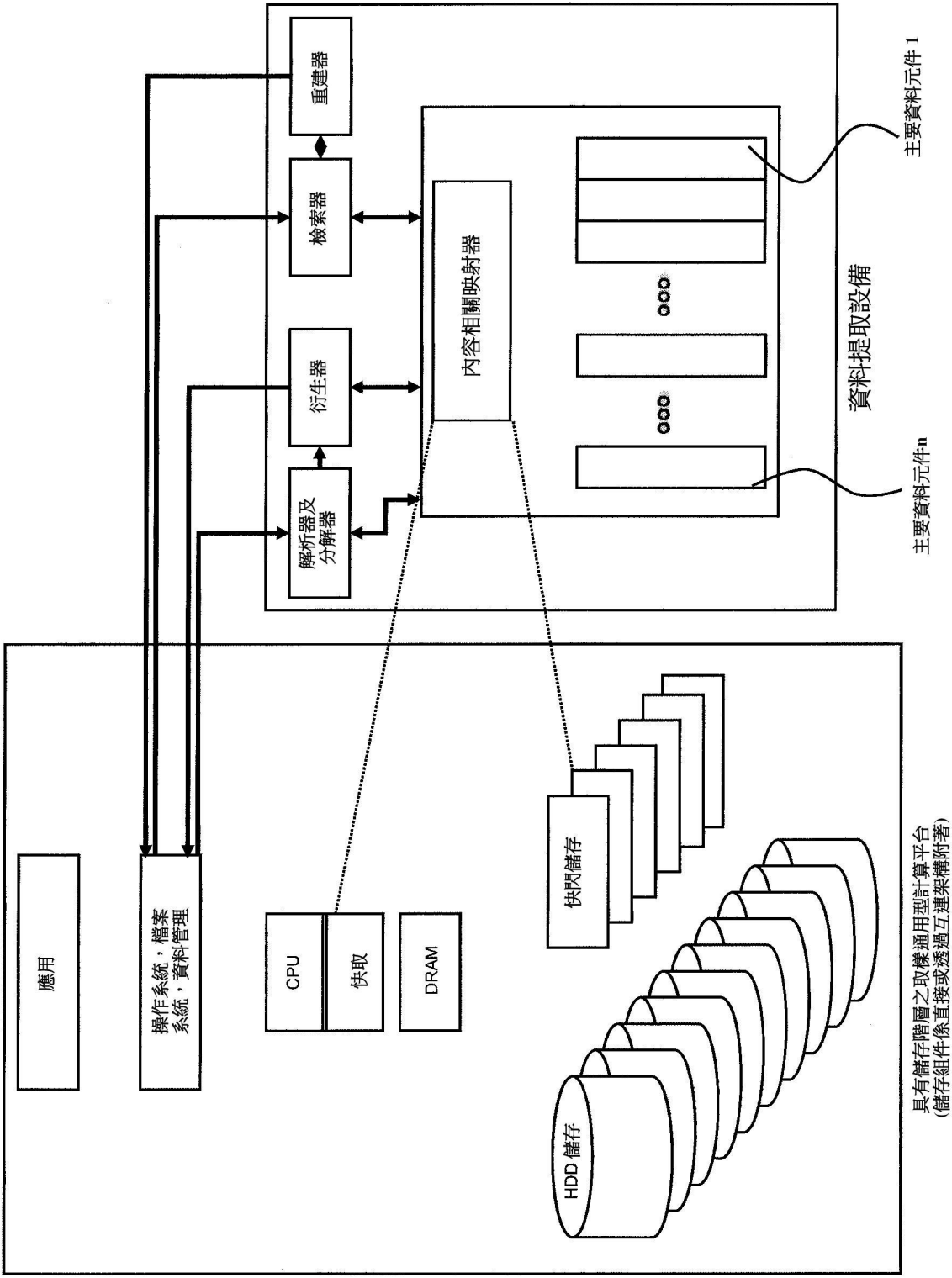
第 11E 圖

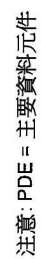


第 11F 圖



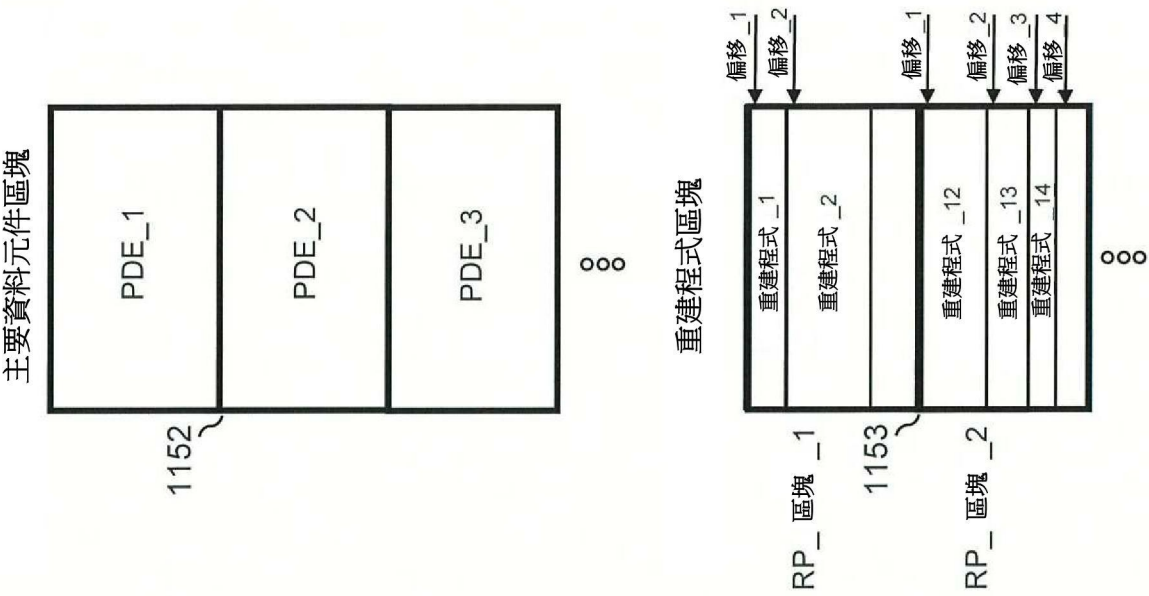
第 11G 圖



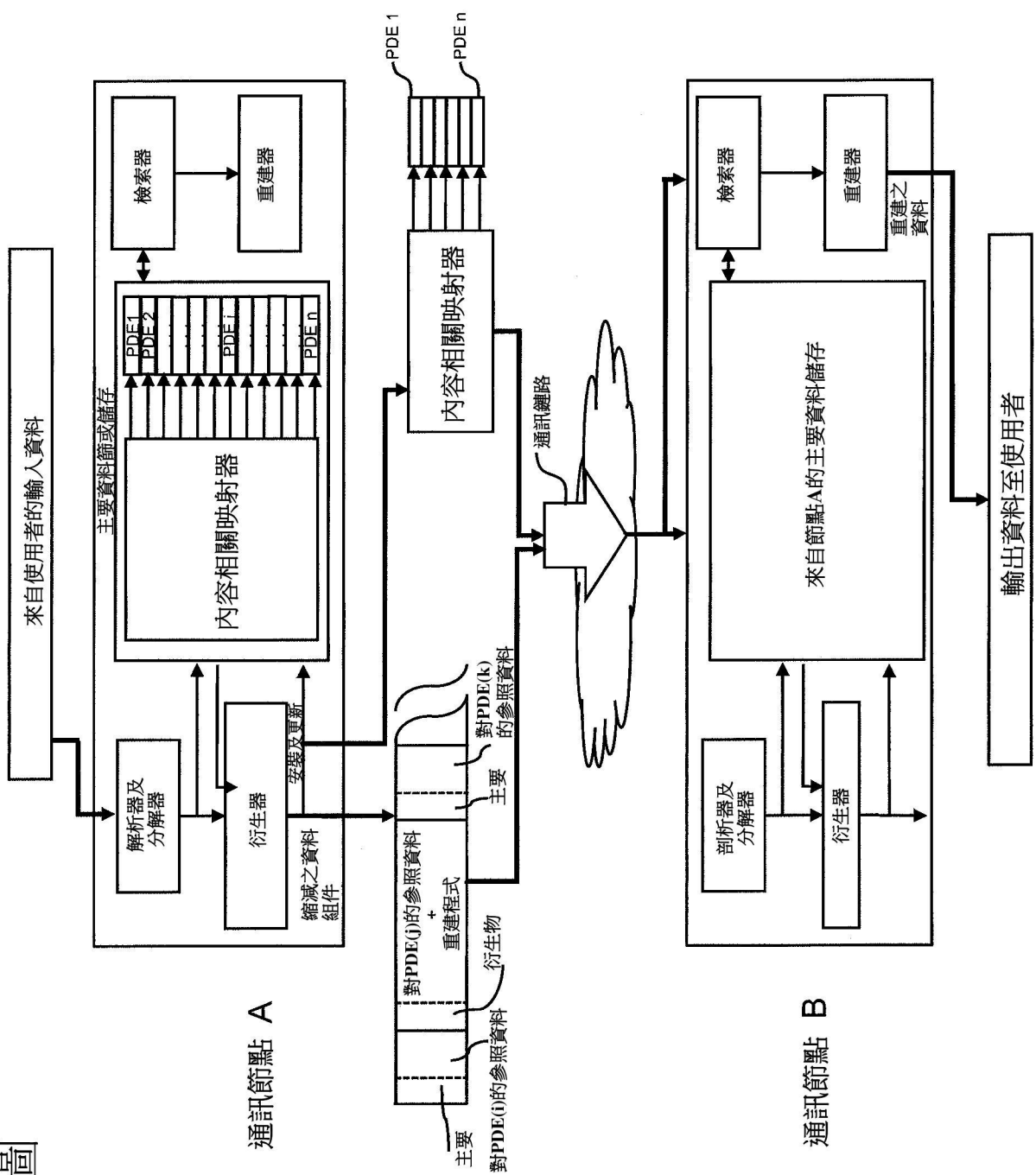


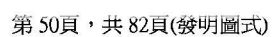
第11I 圖

| 邏輯
區塊
位址
(LBA) | 無損縮減形式的提取資料
或元件規範 | | 對於主要資料元件的
葉節點結構中的計數
欄位的參照 | 對於重建程式的
葉節點結構中的計數
欄位的參照 |
|-------------------------|----------------------|---|---------------------------------|-------------------------------|
| | 元件的
操作碼
或類型 | 參照相關的主要
資料元件區塊及
在重建程式區塊
中的重建程式 | | |
| 1 | 主要的 | PDE_1 | PDE葉節點_7,
計數_4,
計數_PDE_1 | 不適用 |
| 2 | 衍生的 | PDE_1,
RP_區塊_1,
偏移_1 | PDE葉節點_7,
子_4,
計數_PDE_1 | RP葉節點_5,
子_2,
計數_RP_1 |
| 3 | 主要的 | PDE_2 | PDE葉節點_9,
子_12,
計數_PDE_2 | 不適用 |
| 4 | 衍生的 | PDE_1,
RP_區塊_1,
偏移_2 | PDE葉節點_7,
子_4,
計數_PDE_1 | RP葉節點_5,
子_2,
計數_RP_2 |
| ... | ... | ... | ... | ... |
| 19 | 衍生的 | PDE_3,
RP_區塊_2,
偏移_3 | PDE葉節點_14,
子_6,
計數_PDE_3 | RP葉節點_5,
子_2,
計數_RP_14 |
| ... | ... | ... | ... | ... |

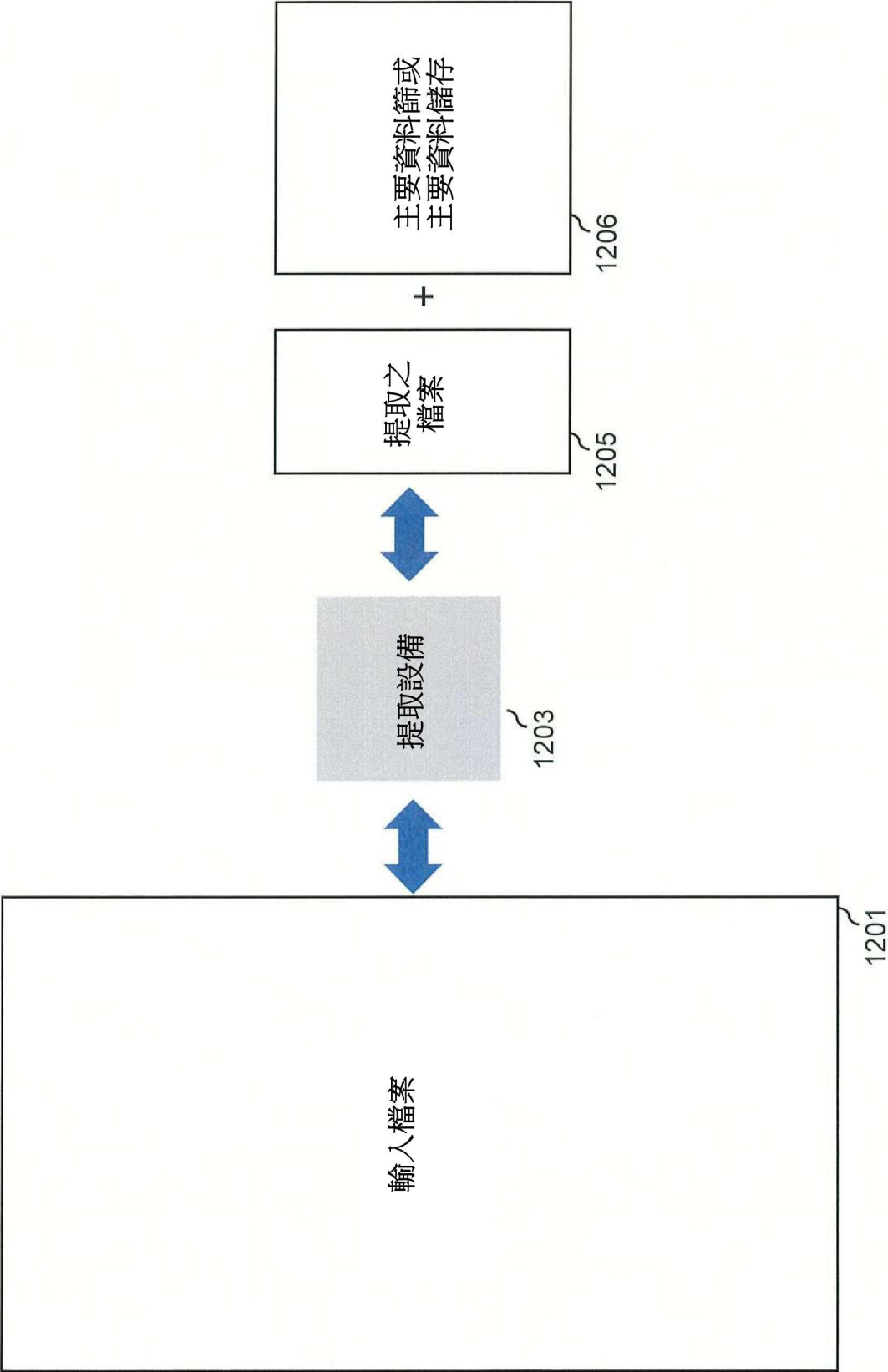


第 12A 圖

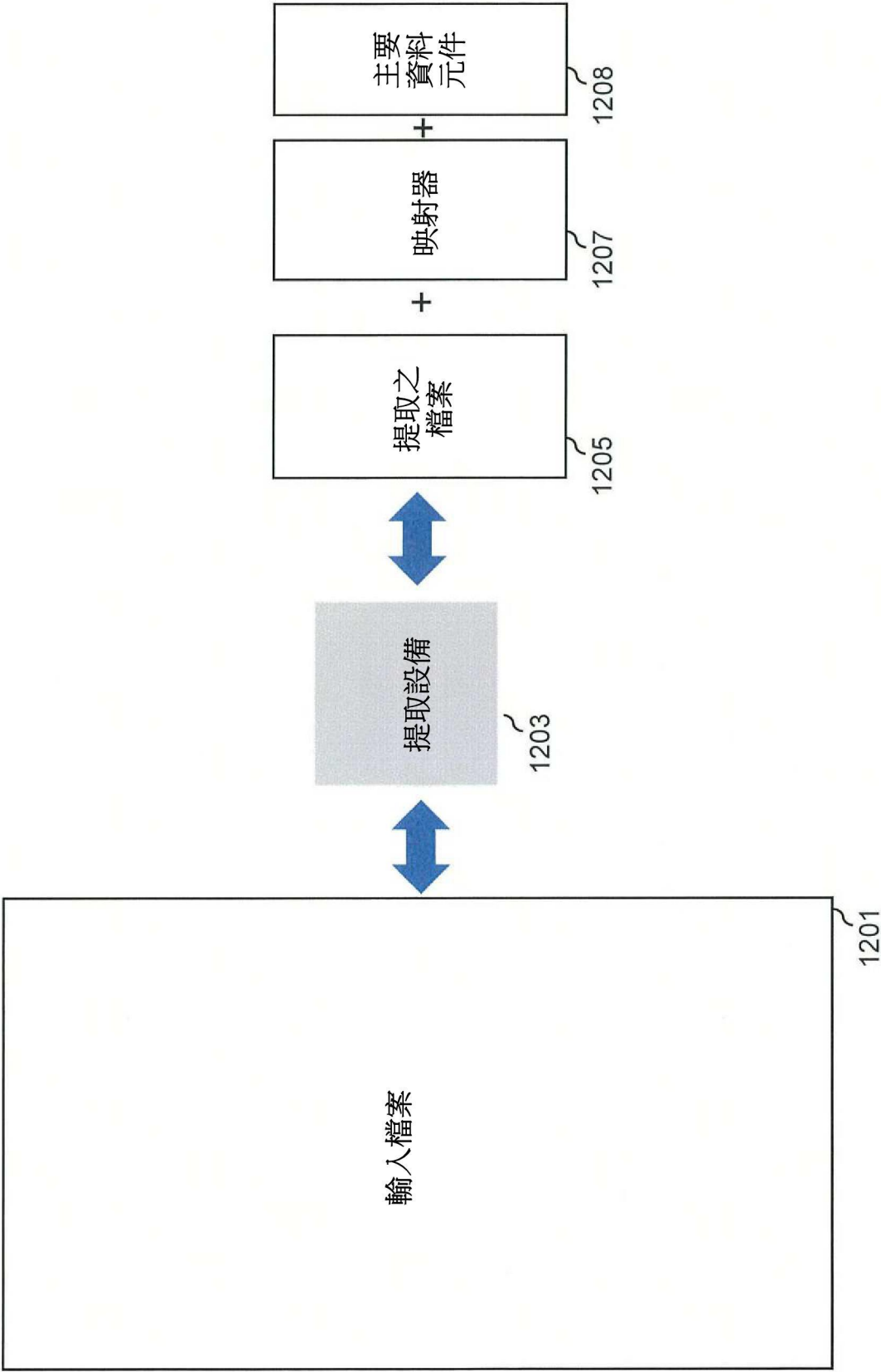




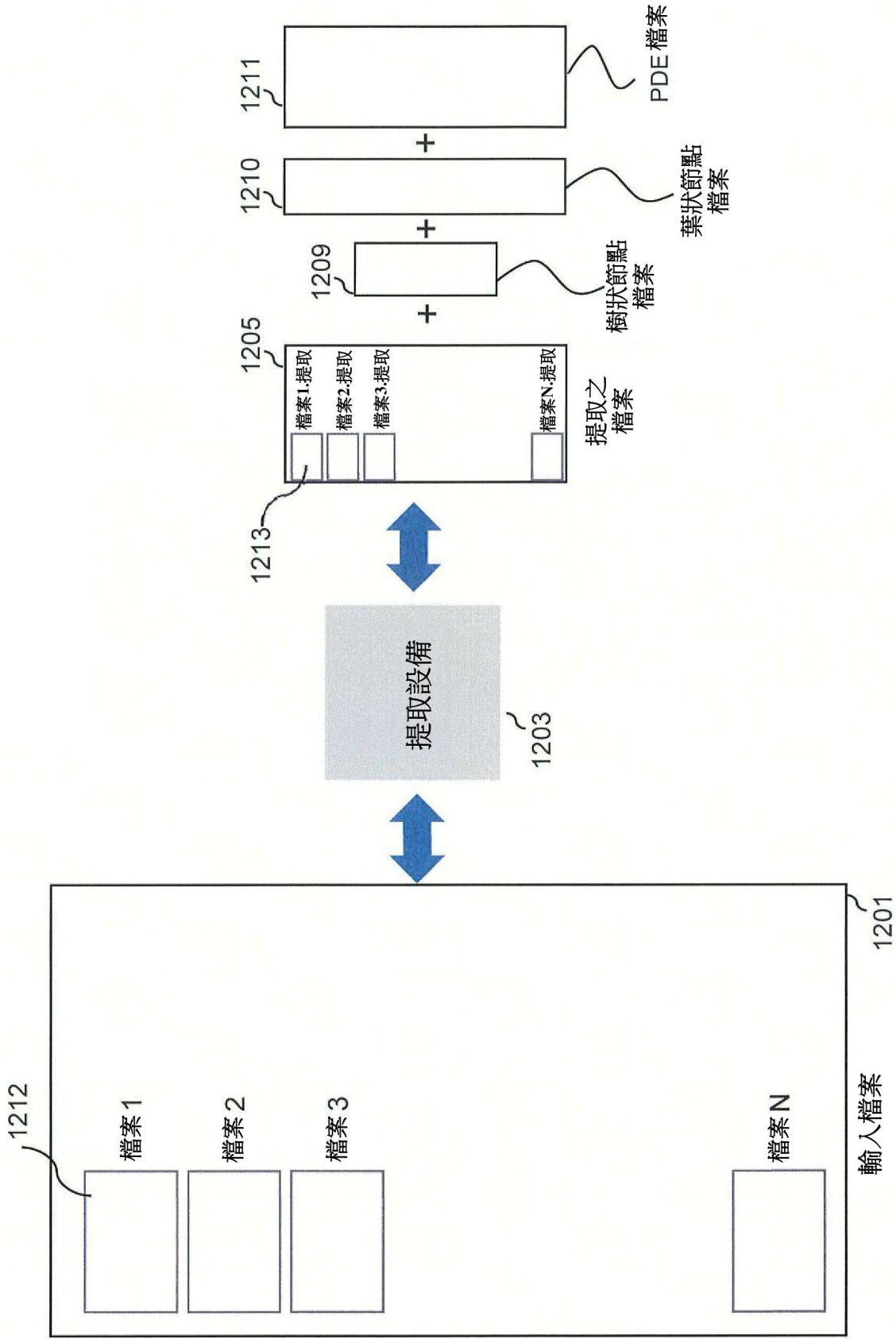
第 12C 圖



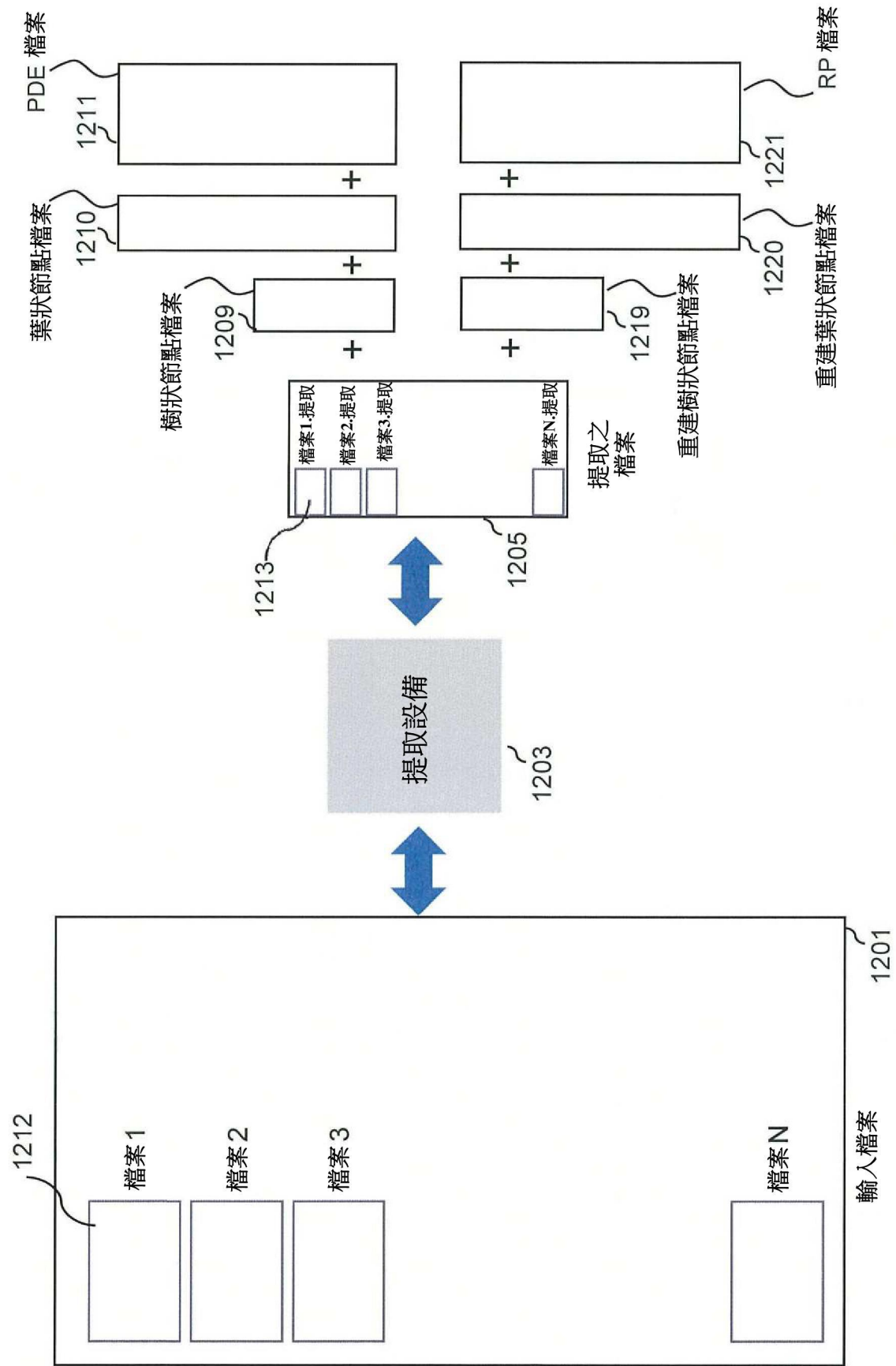
第12D圖



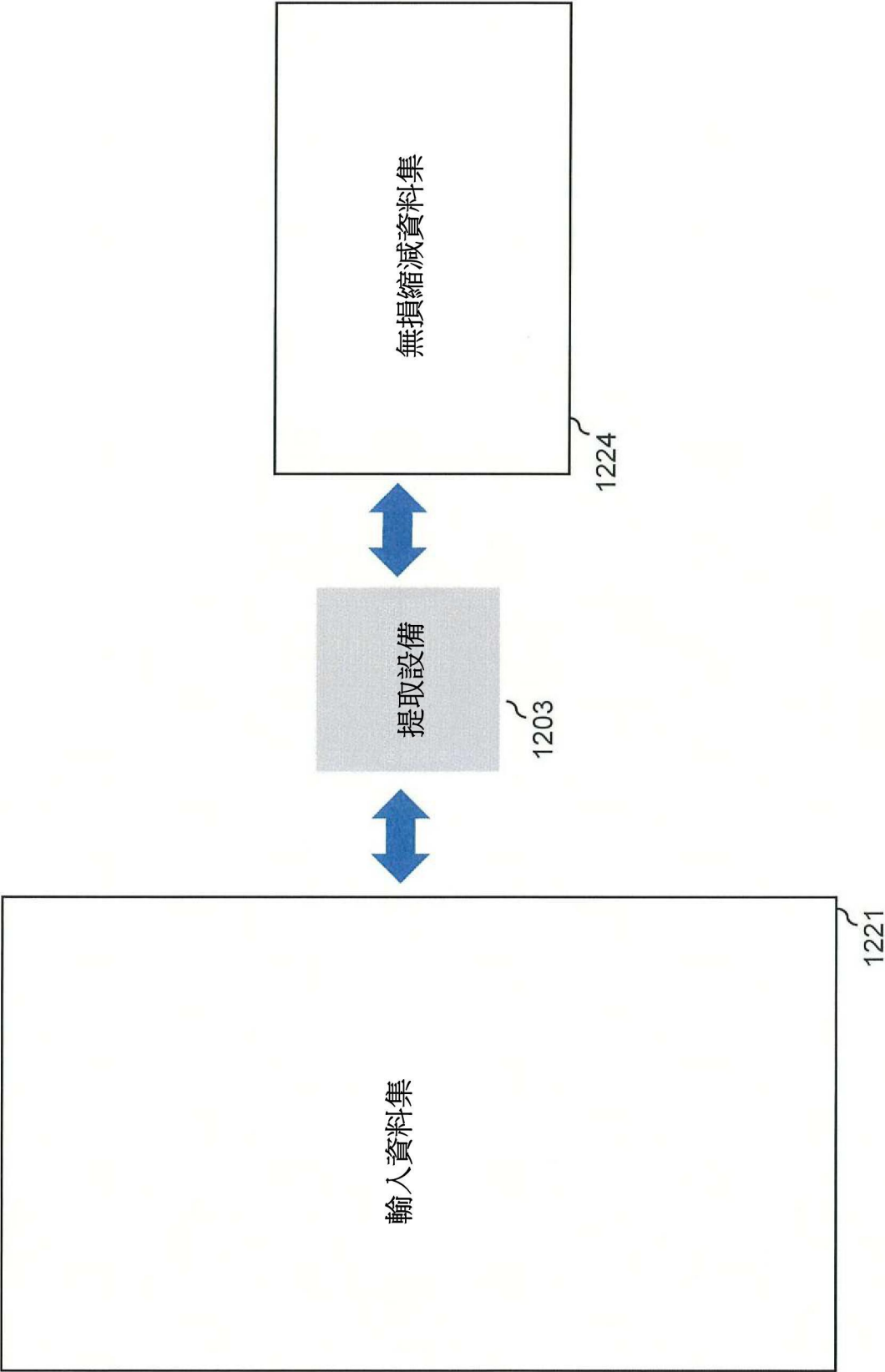
第 12E 圖



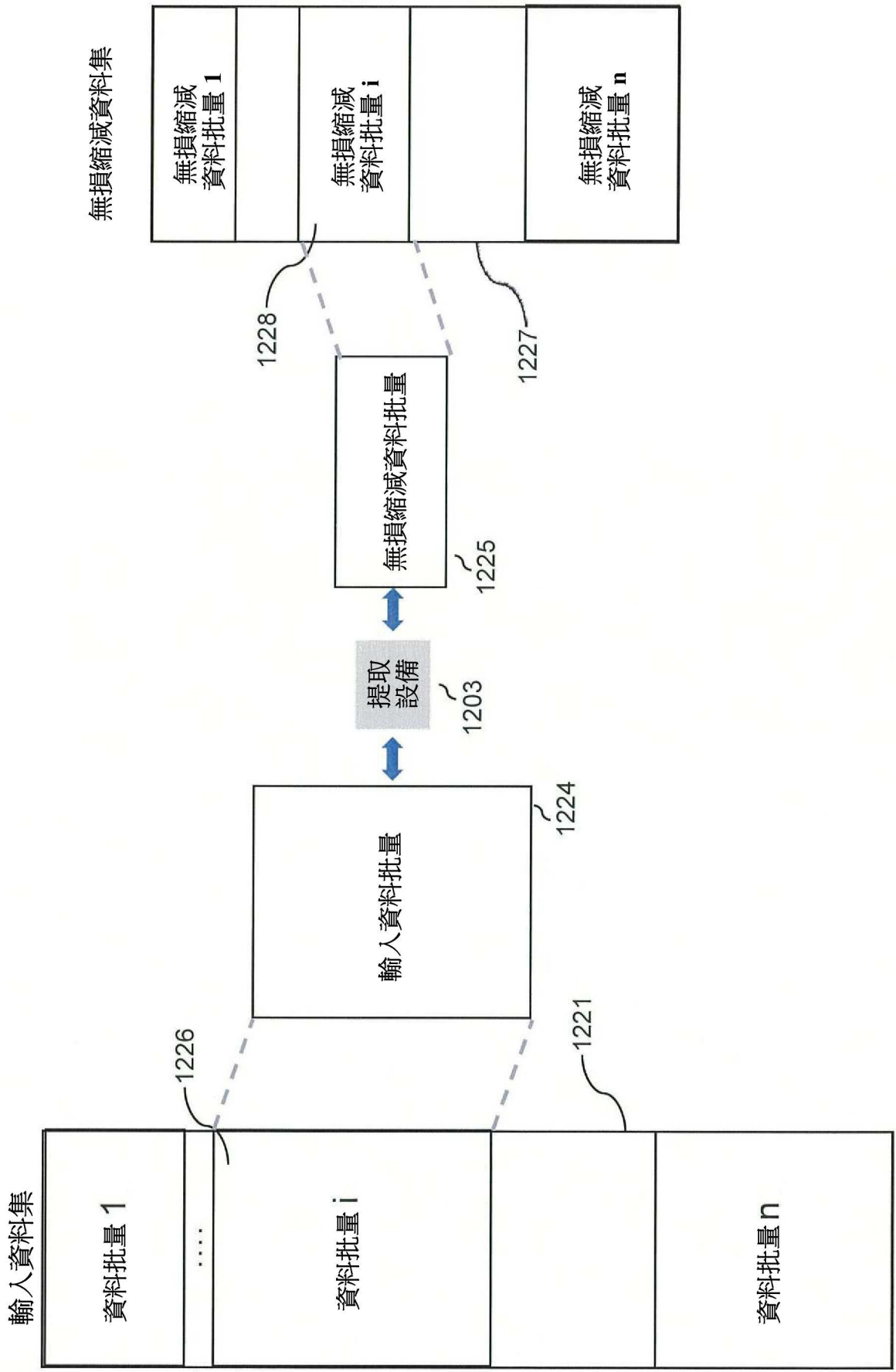
第 12F 圖



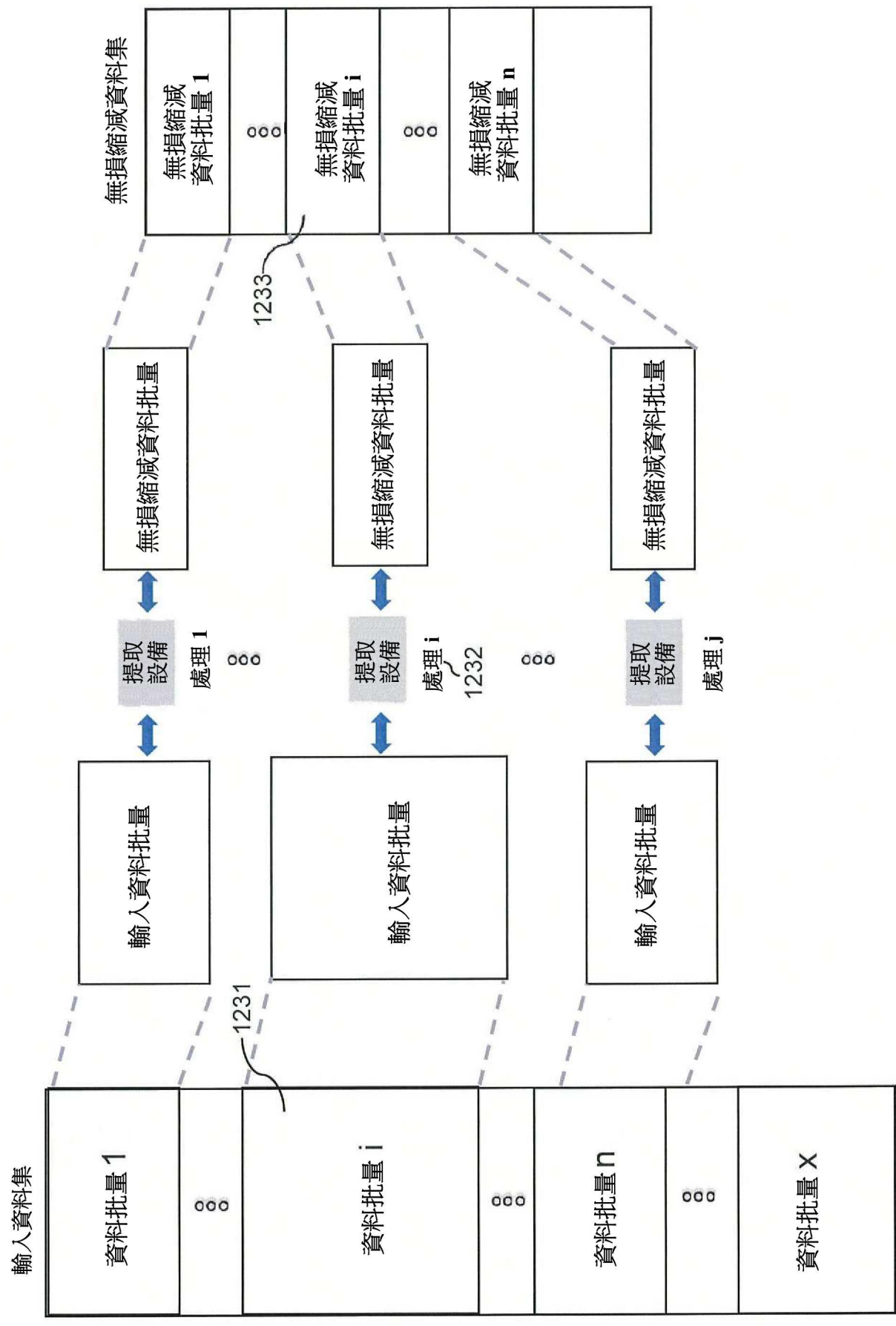
第 12G 圖



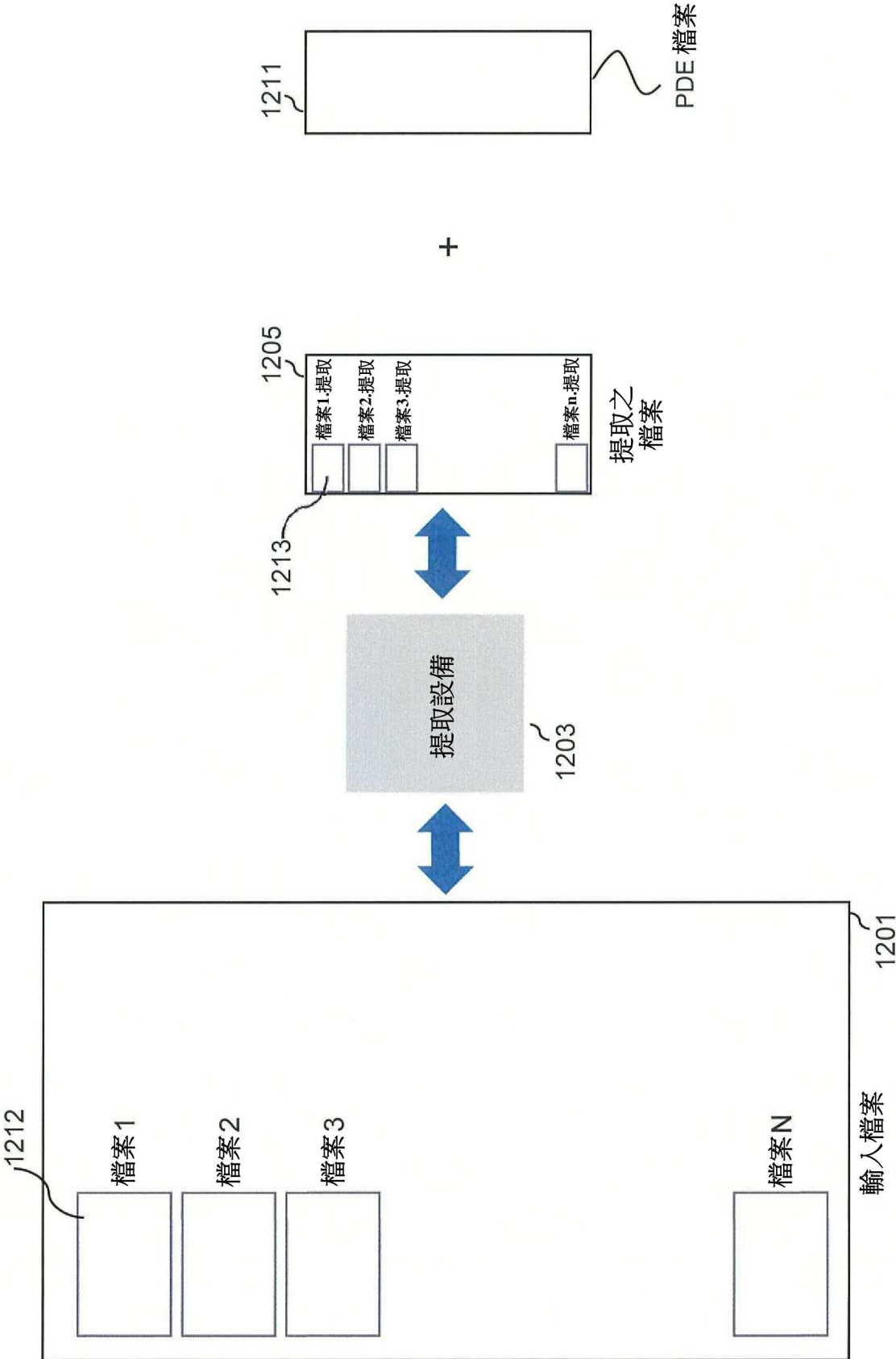
第 12H 圖



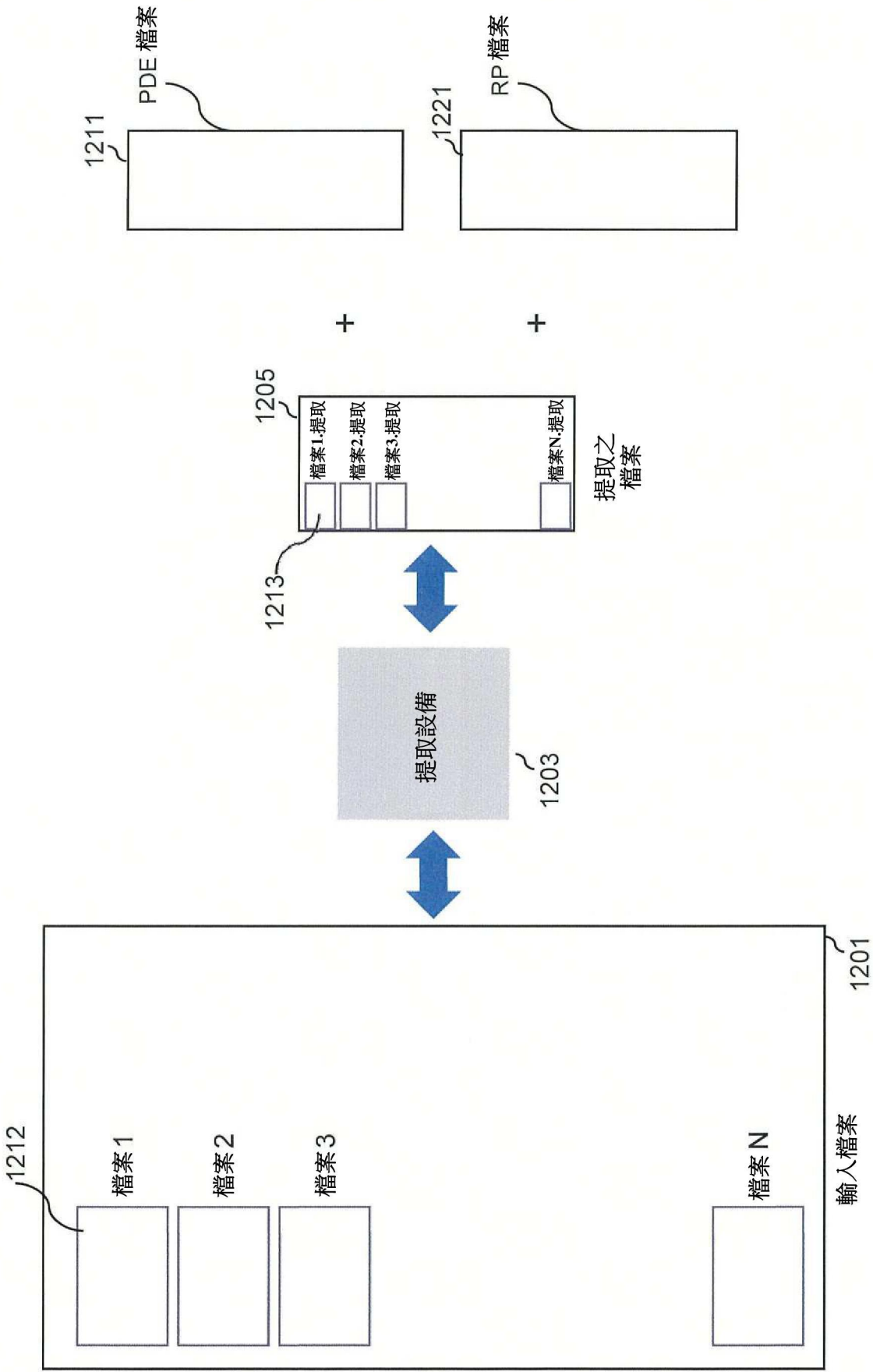
第 12I 圖



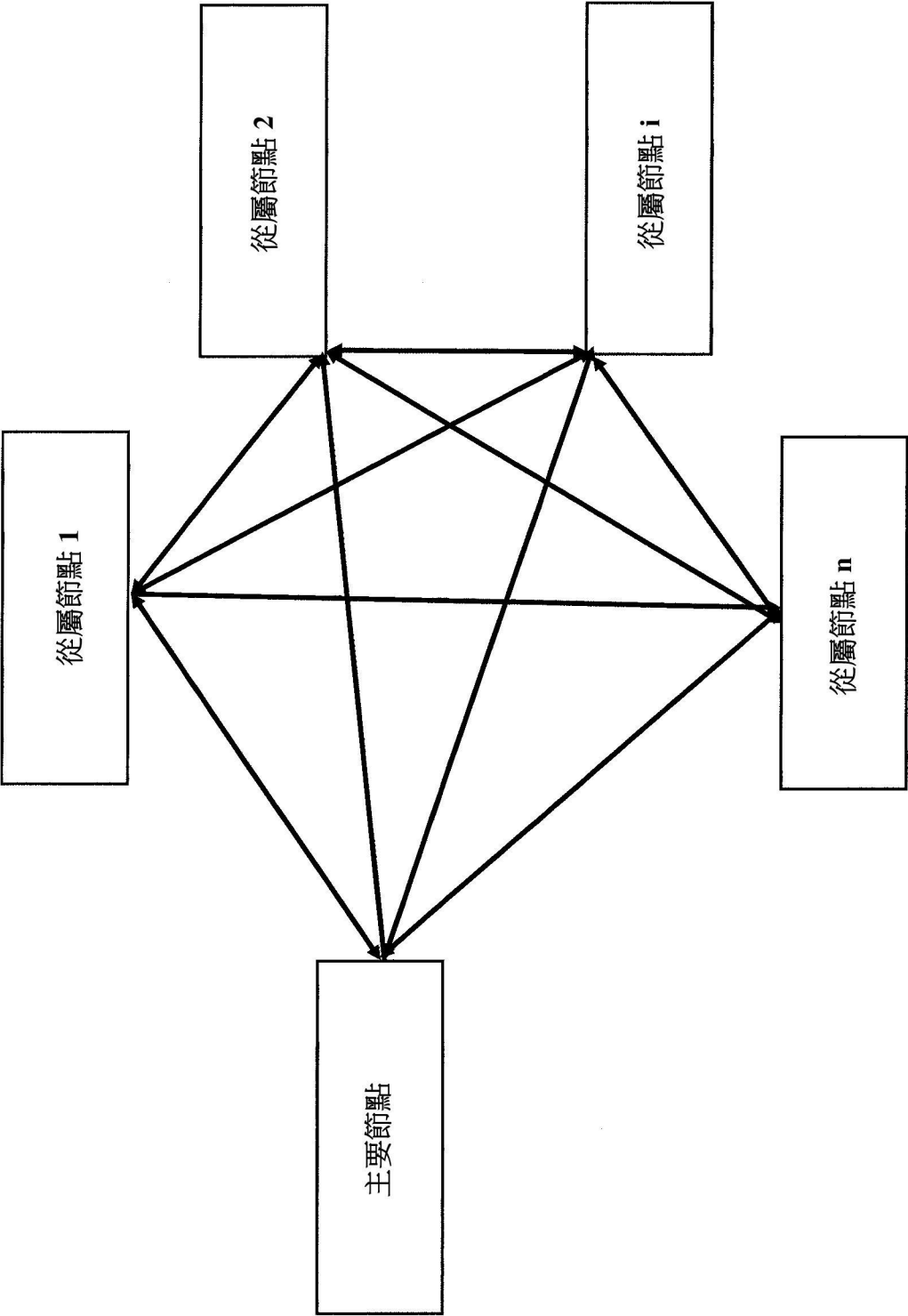
第 12J 圖



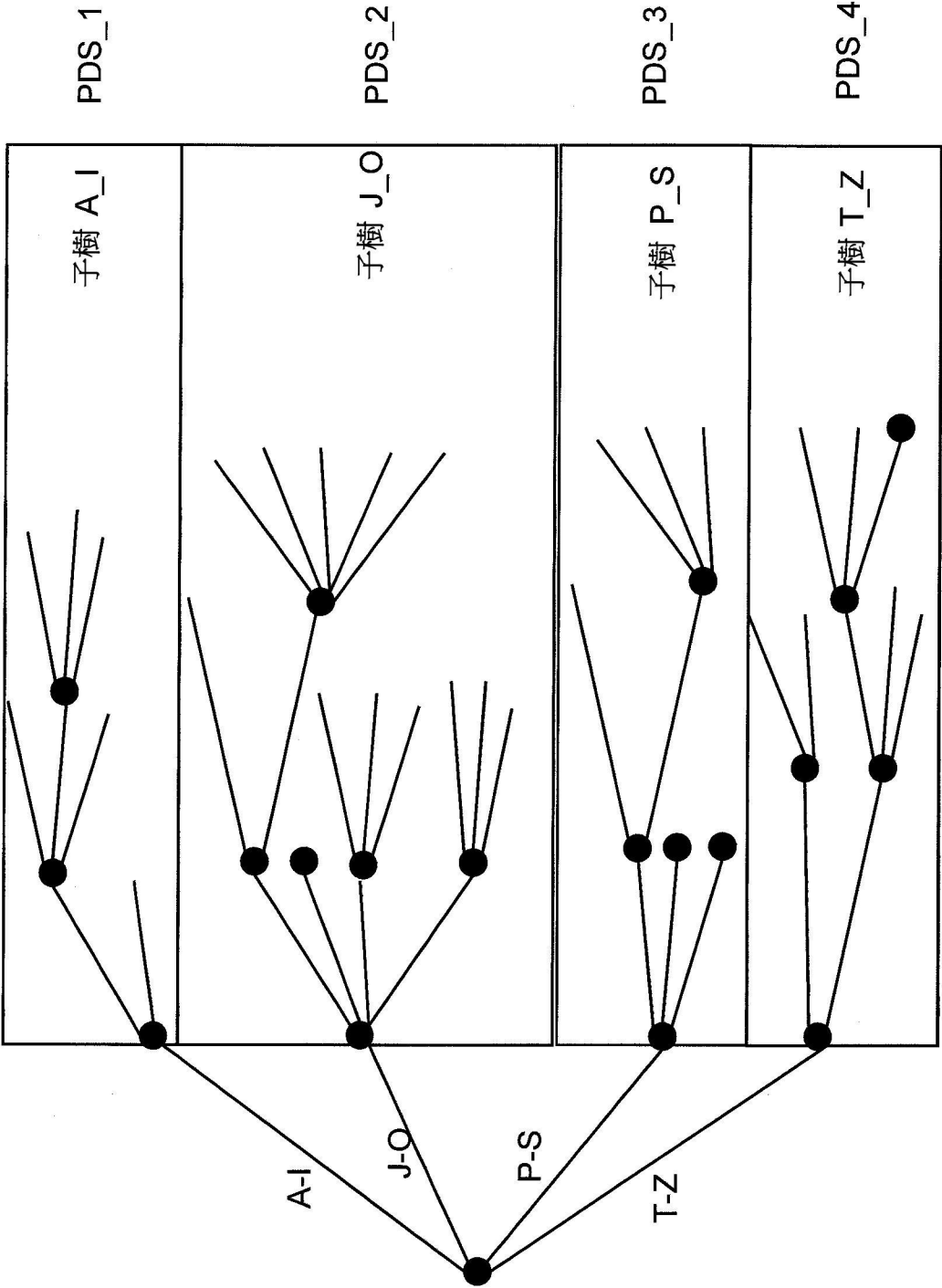
第 12K 圖



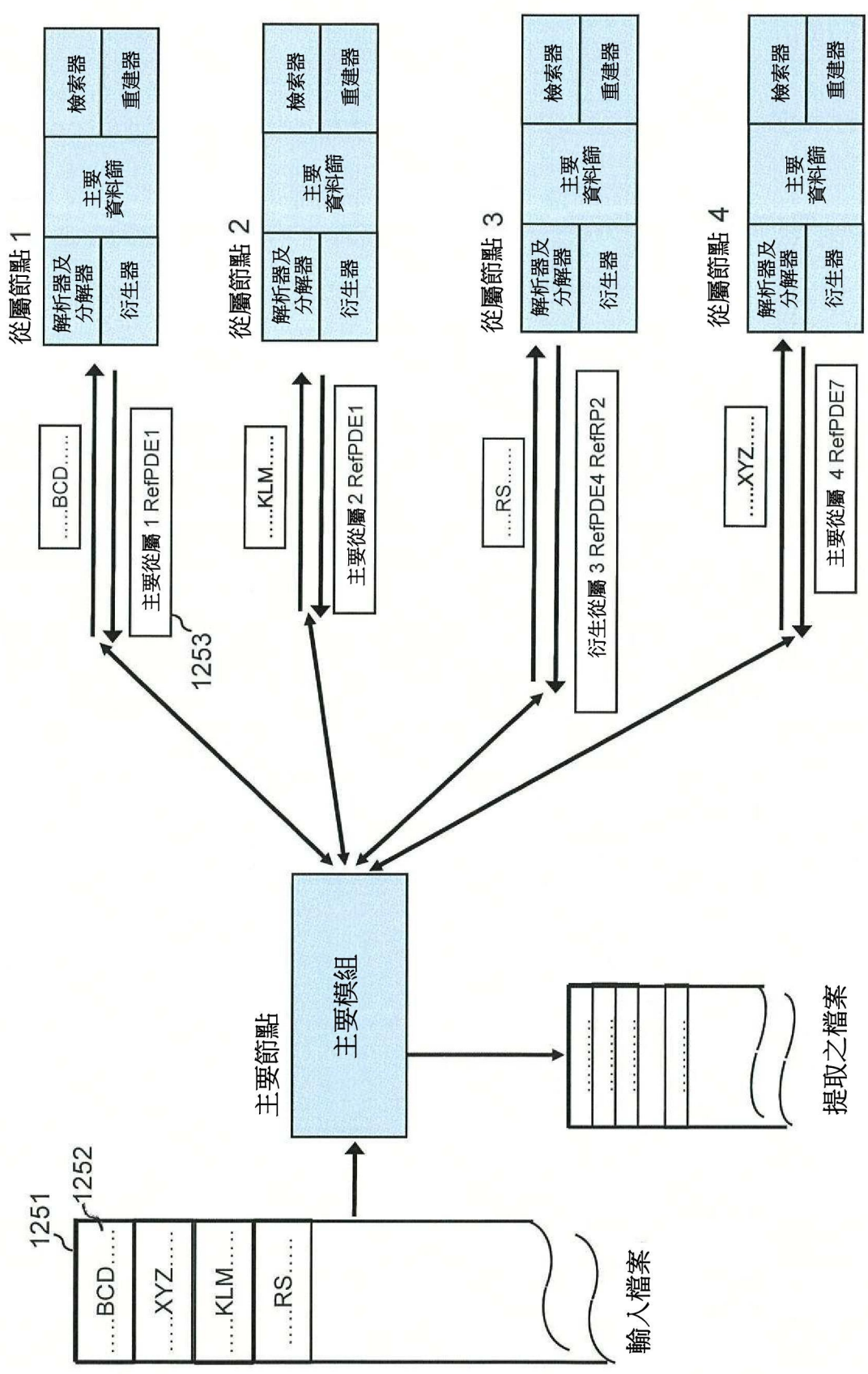
第 12L 圖



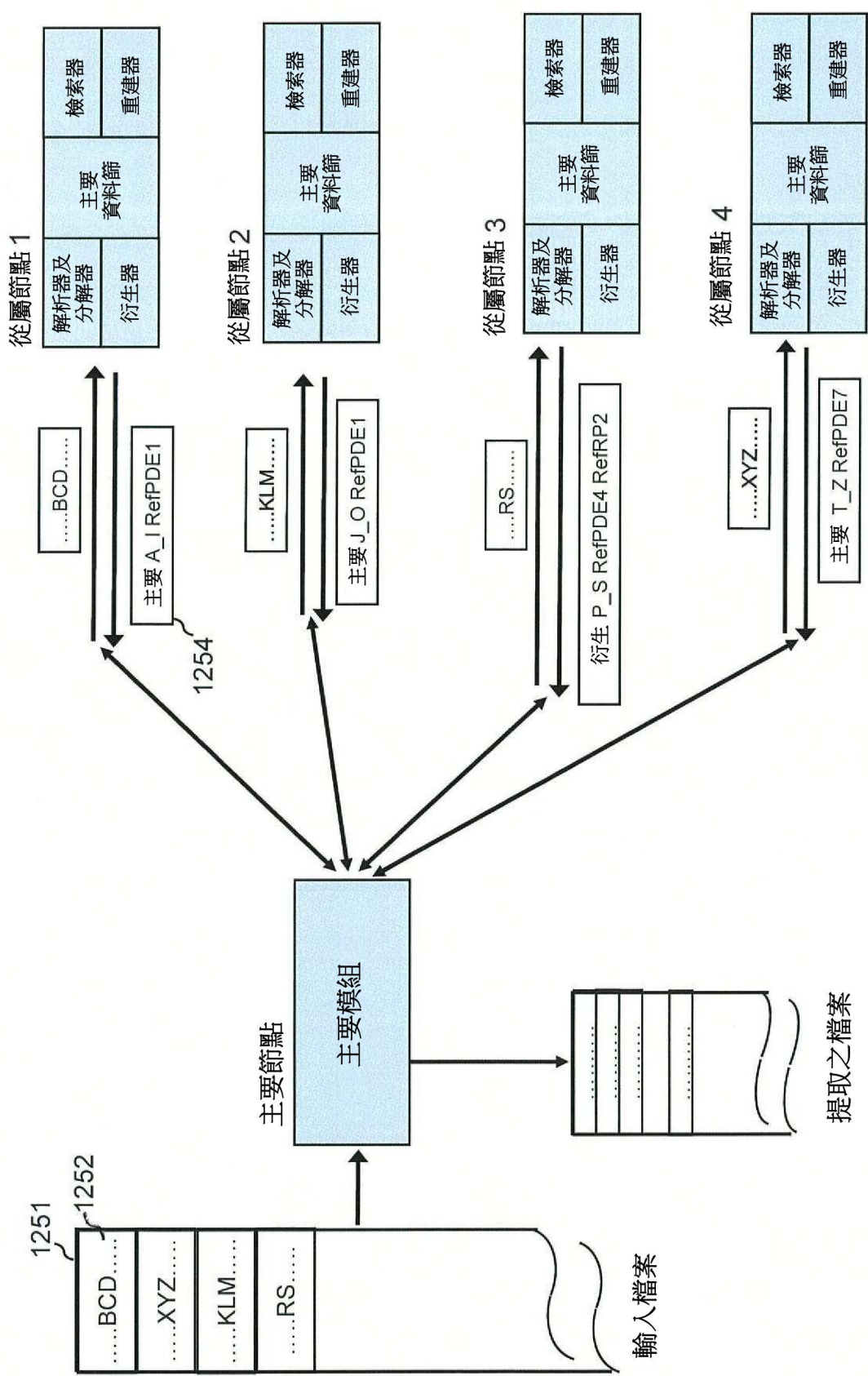
第 12M 圖



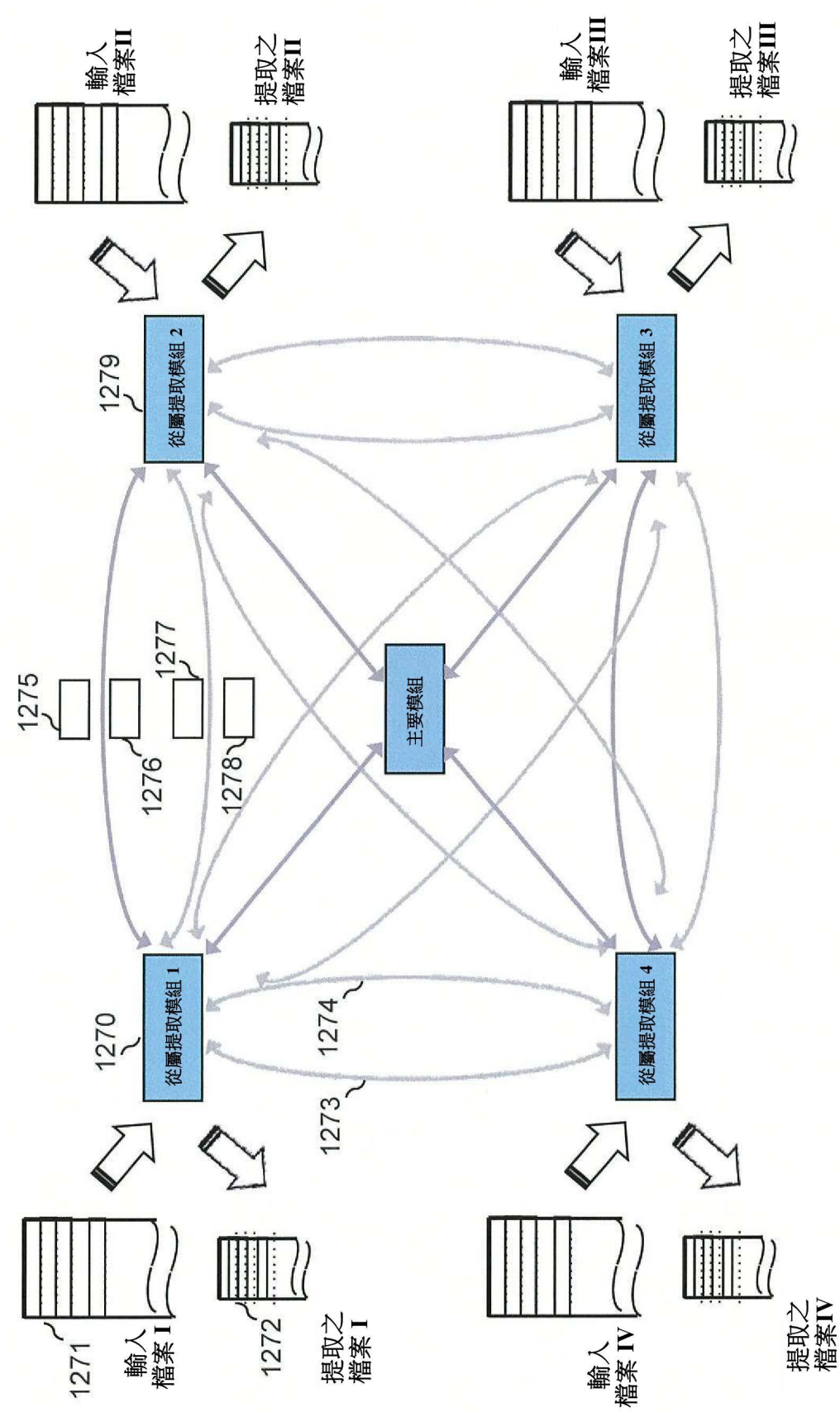
第 12N 圖



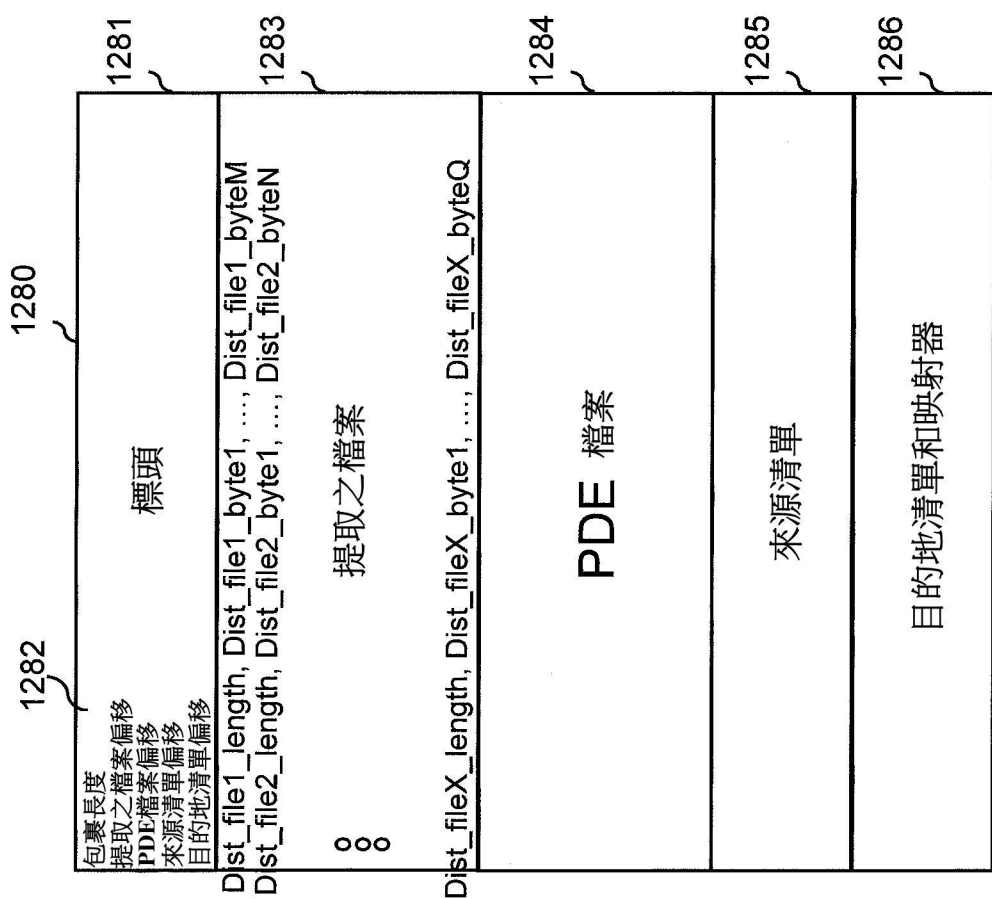
第 120 圖



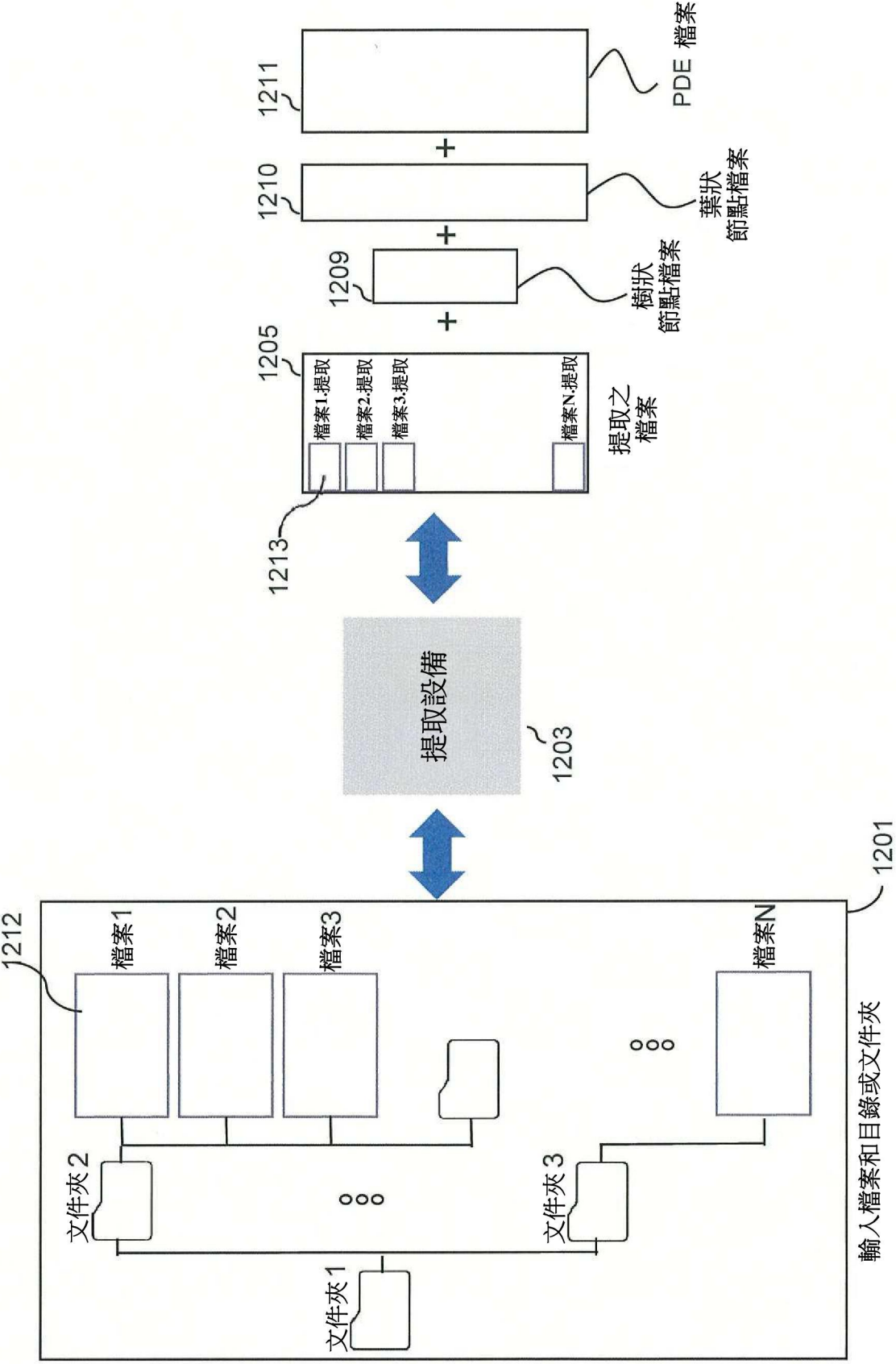
第12P圖



第 12Q 圖



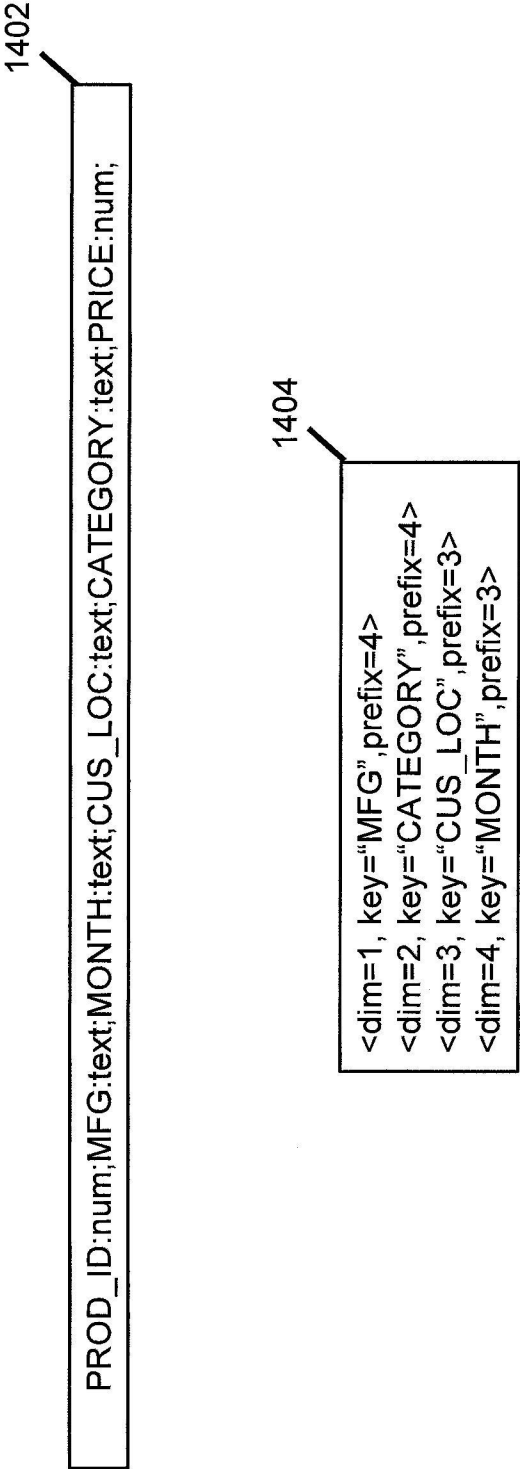
第 12R 圖



第 13 圖

| 路徑資訊 | | | | | | | | |
|------|----------|----------|-----------------|----------|----------------|------------------|---|--|
| 子的數目 | | | | | | | | |
| N | | | | | | | | |
| 子ID | 不同位元組的數目 | 不同位元組的值 | 對於主要資料
元件的參照 | 導航先行位元組 | 複製品和
衍生物的計數 | 主要資料元件
的其他元資料 | 對於在提取資料中的
每個元件的反向參照
以及元件是否為主要
的或衍生的之指示符 | |
| 1 | 1 | 17 | pde 76 | 13476231 | 4 | <...> | Elem 25: prime
Elem 43: prime
Elem 47: deriv
..... | |
| 2 | 1 | 32 | pde 4718 | 00337650 | 7 | <...> | Elem 62: prime
..... | |
| ... | ... | ... | ... | ... | ... | ... | | |
| N | 4 | 654aed21 | pde 786 | ed189721 | 12 | <...> | Elem 13: prime
Elem 96: deriv
Elem 97: deriv
Elem 98: prime
Elem 99: deriv
..... | |

第 14A 圖



第14B圖

檔案

1405

| | |
|-------|---|
| 交易 1 | PROD_ID:616;MFG:PRINCE;MONTH:FEB;CUS_LOC:NYC;CATEGORY:RACQ;PRICE:\$190.69 ... |
| 交易 2 | PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$170.80 ... |
| 交易 3 | PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$160.80 ... |
| 交易 4 | PROD_ID:251;MFG:WILSON;MONTH:NOV;CUS_LOC:LAH;CATEGORY:BALL;PRICE:\$27.50 ... |
| 交易 58 | PROD_ID:349;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$125.32 ... |
| 交易 59 | PROD_ID:427;MFG:NIKE;MONTH:OCT;CUS_LOC:LAH;CATEGORY:JERSEY;PRICE:\$75.95 ... |
| 交易 60 | PROD_ID:699;MFG:ADIDAS;MONTH:SEP;CUS_LOC:SJC;CATEGORY:SHOE;PRICE:\$917.25 ... |



1406

1407

候選元件 1, D1 D2 D3 D4=PRIN RACQ NYC FEB
候選元件 2, D1 D2 D3 D4=NIKE SHOE LAH JUN
候選元件 3, D1 D2 D3 D4=NIKE SHOE LAH JUN
候選元件 4, D1 D2 D3 D4=WILS BALL LAH NOV

候選元件 58, D1 D2 D3 D4=NIKE SHOE LAH JUN
候選元件 59, D1 D2 D3 D4=NIKE JERS LAH OCT
候選元件 60, D1 D2 D3 D4=ADID SHOE SJC SEP

| |
|---|
| PROD_ID:616;MFG:PRINCE;MONTH:FEB;CUS_LOC:NYC;CATEGORY:RACQ;PRICE:\$190.69 ... |
| PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$170.80 ... |
| PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$160.80 ... |
| PROD_ID:251;MFG:WILSON;MONTH:NOV;CUS_LOC:LAH;CATEGORY:BALL;PRICE:\$27.50 ... |
| PROD_ID:349;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$125.32 ... |
| PROD_ID:427;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:JERS;PRICE:\$75.95 ... |
| PROD_ID:699;MFG:ADIDAS;MONTH:SEP;CUS_LOC:SJC;CATEGORY:SHOE;PRICE:\$917.25 ... |

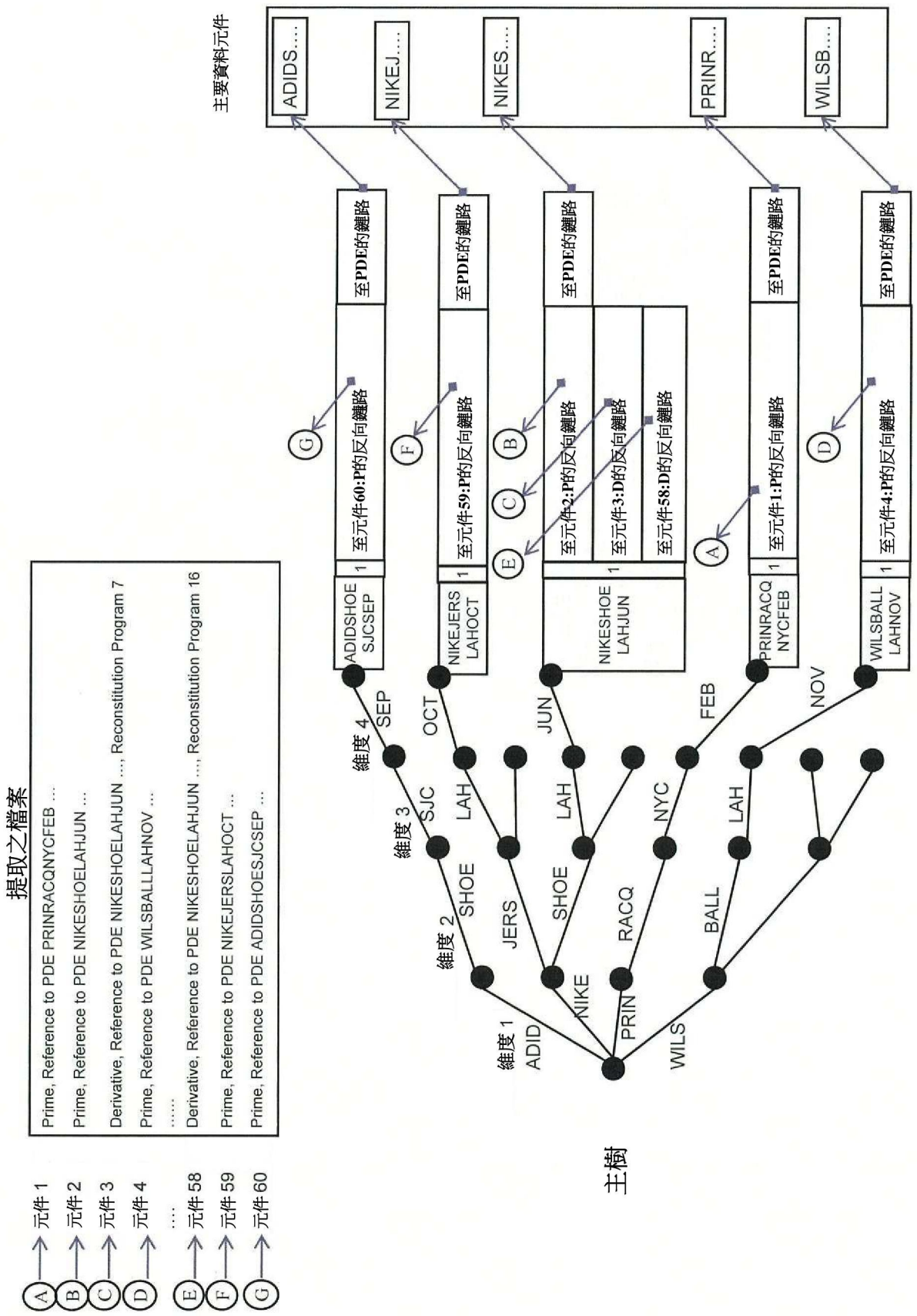


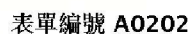
1408

提取之檔案

| | |
|-------|--|
| 元件 1 | Prime, Reference to PDE PRINRACQNYCFEB ... |
| 元件 2 | Prime, Reference to PDE NIKESHOELAHJUN ... |
| 元件 3 | Derivative, Reference to PDE NIKESHOELAHJUN ..., Reconstitution Program 7 |
| 元件 4 | Prime, Reference to PDE WILSBALLAHNOV ... |
| 元件 58 | Derivative, Reference to PDE NIKESHOELAHJUN ..., Reconstitution Program 16 |
| 元件 59 | Prime, Reference to PDE NIKEJERSLAHOCT ... |
| 元件 60 | Prime, Reference to PDE ADIDSHOESJCSEP ... |

第14C 圖

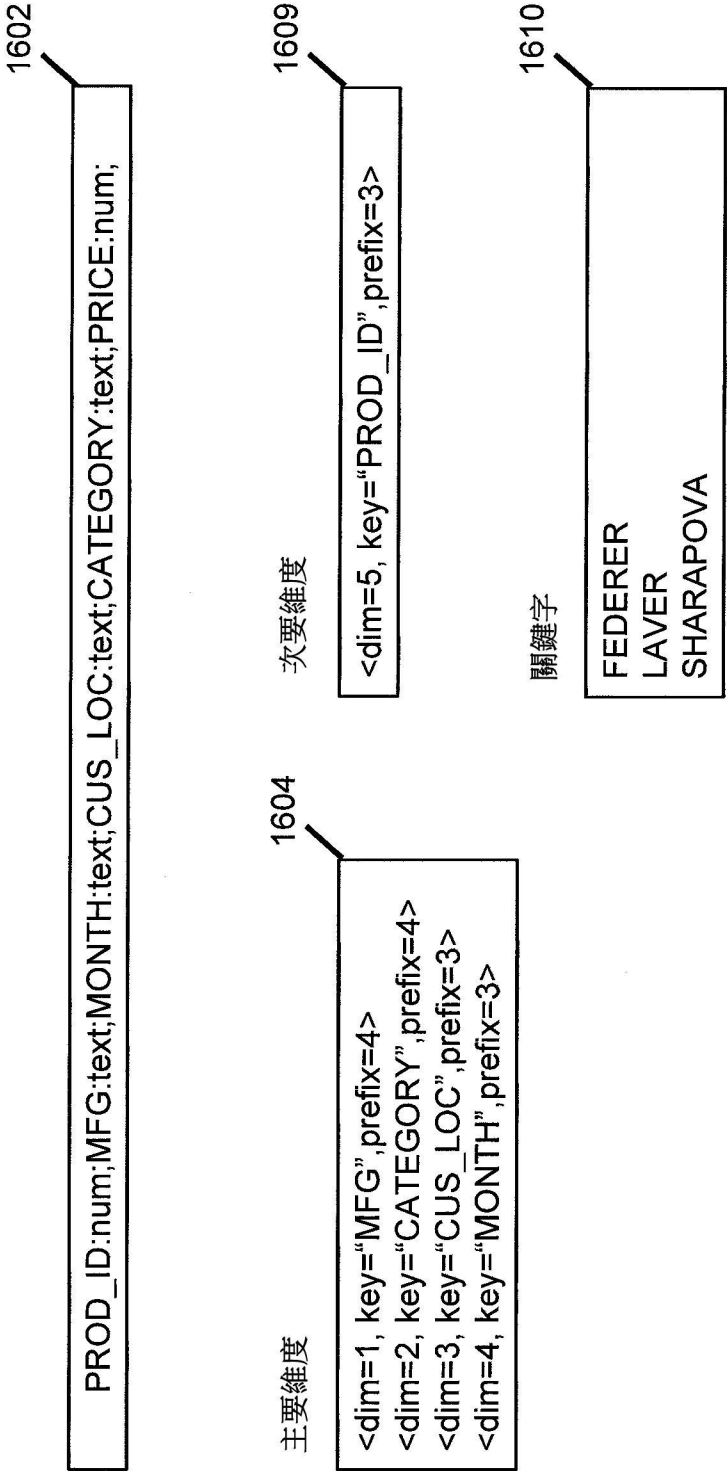




第 15 圖

| 關鍵字 | 對於提取檔案中的
元件的反向參照
(檔案名稱, 元件#) | 輸入資料中的重建
元件中的偏移 |
|-------|--------------------------------------|--------------------|
| 關鍵字 1 | File_FOO, Elem 11 | 90 |
| | File_XYZ, Elem 251 | 3043 |
| | File_ABC, Elem 9833 | 682 |
| 關鍵字 2 | File_Misc, Elem 61773 | 2080 |
| 關鍵字 3 | File_MyName, Elem 3 | 1641 |
| | File_XYZ, Elem 405 | 808 |
| ... | ... | ... |
| 關鍵字 N | File FOO, Elem 2002 | 1740 |

第 16A 圖



第16B圖

檔案 1611

| | |
|-------|--|
| 交易 1 | PROD_ID:616;MFG:PRINCE;MONTH:FEB;CUS_LOC:NYC;CATEGORY:RACQ;PRICE:\$190.69..SHARAPOVA |
| 交易 2 | PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$170.80..FEDERER |
| 交易 3 | PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$160.80..FEDERER |
| 交易 4 | PROD_ID:251;MFG:WILSON;MONTH:NOV;CUS_LOC:LAH;CATEGORY:BALL;PRICE:\$27.50.. |
| 交易 58 | PROD_ID:349;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$125.32..FEDERER |
| 交易 59 | PROD_ID:427;MFG:NIKE;MONTH:OCT;CUS_LOC:LAH;CATEGORY:JERSEY;PRICE:\$75.95 ... |
| 交易 60 | PROD_ID:307;MFG:ADIDAS;MONTH:SEP;CUS_LOC:SJC;CATEGORY:SHOE;PRICE:\$917.25..LAVER |

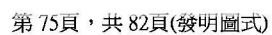
維度：
主要維度 (D1)=MFG, 前4字元
主要維度 (D2)=CATEGORY, 前4字元
主要維度 (D3)=CUS\_LOC, 前3字元
主要維度 (D4)=MONTH, 前3字元
次要維度 = PROD\_ID, 前3字元

1612 1613

| | |
|--|---|
| 候選元件 1, D1 D2 D3 D4=PRIN RACQ NYC FEB | PROD_ID:616;MFG:PRINCE;MONTH:FEB;CUS_LOC:NYC;CATEGORY:RACQ;PRICE:\$190.69 ... |
| 候選元件 2, D1 D2 D3 D4=NIKE SHOE LAH JUN | PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$170.80 ... |
| 候選元件 3, D1 D2 D3 D4=NIKE SHOE LAH JUN | PROD_ID:348;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$160.80 ... |
| 候選元件 4, D1 D2 D3 D4=WILS BALL LAH NOV | PROD_ID:251;MFG:WILSON;MONTH:NOV;CUS_LOC:LAH;CATEGORY:BALL;PRICE:\$27.50 ... |
| 候選元件 58, D1 D2 D3 D4=NIKE SHOE LAH JUN | PROD_ID:349;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:SHOE;PRICE:\$125.32 ... |
| 候選元件 59, D1 D2 D3 D4=NIKE JERS LAH OCT | PROD_ID:427;MFG:NIKE;MONTH:JUN;CUS_LOC:LAH;CATEGORY:JERS;PRICE:\$75.95 ... |
| 候選元件 60, D1 D2 D3 D4=ADID SHOE SJC SEP | PROD_ID:307;MFG:ADIDAS;MONTH:SEP;CUS_LOC:SJC;CATEGORY:SHOE;PRICE:\$917.25 ... |

提取的檔案 1614

| | |
|-------|---|
| 元件 1 | Prime, Reference to PDE PRINRACQNYCFEB ... |
| 元件 2 | Prime, Reference to PDE NIKESHOELAHJUN ... |
| 元件 3 | Derivative, Reference to PDE NIKESHOELAHJUN ..., Reconstitution Program 7 |
| 元件 4 | Prime, Reference to PDE WILSBALLAHNOV ... |
| 元件 58 | Prime, Reference to PDE NIKESHOELAHJUN ... |
| 元件 59 | Prime, Reference to PDE NIKEJERSLAHOCT ... |
| 元件 60 | Prime, Reference to PDE ADIDSHOESJCSEP ... |

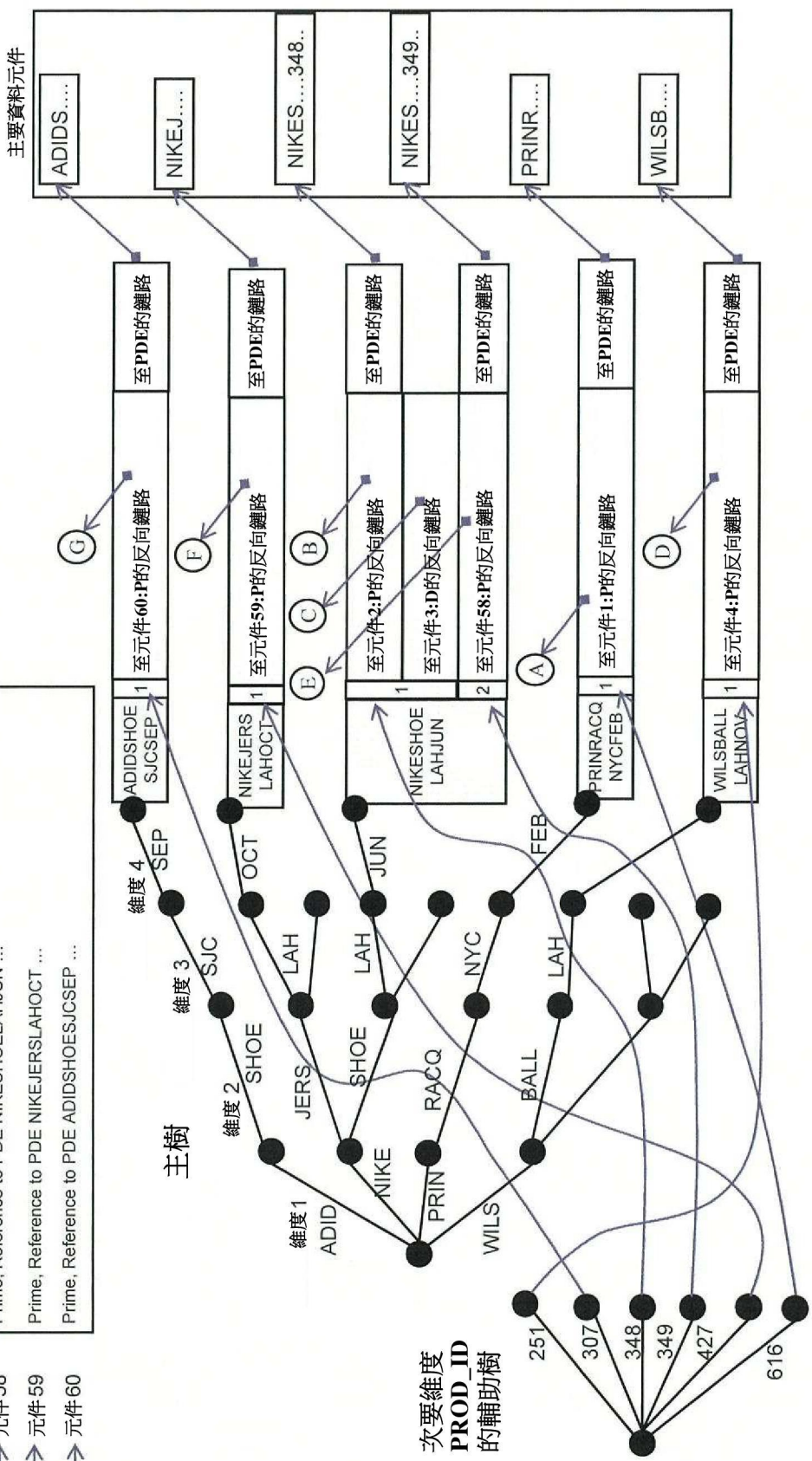


第16D 圖

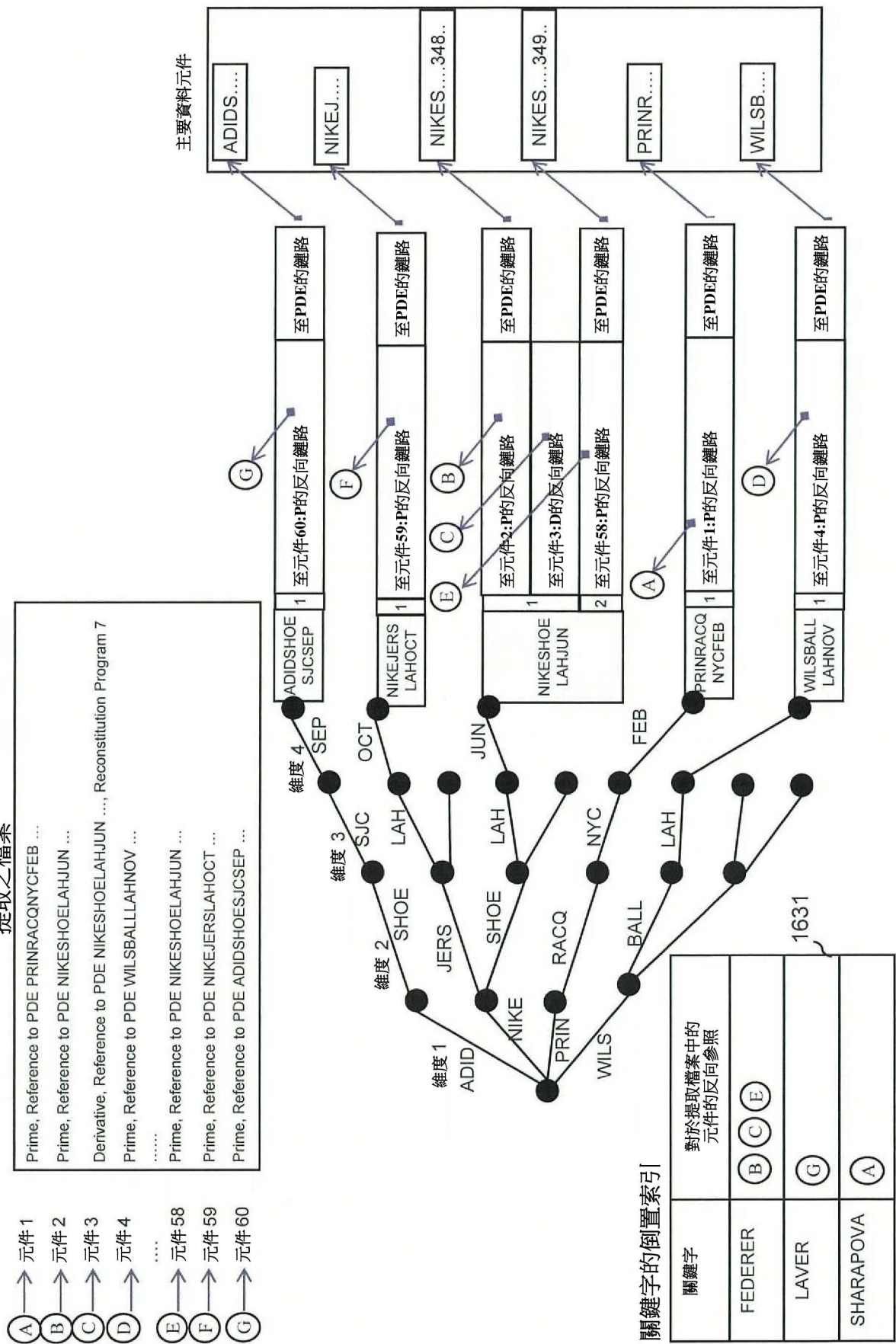
提取之檔案

- A → 元件 1
- B → 元件 2
- C → 元件 3
- D → 元件 4
-
- E → 元件 58
- F → 元件 59
- G → 元件 60

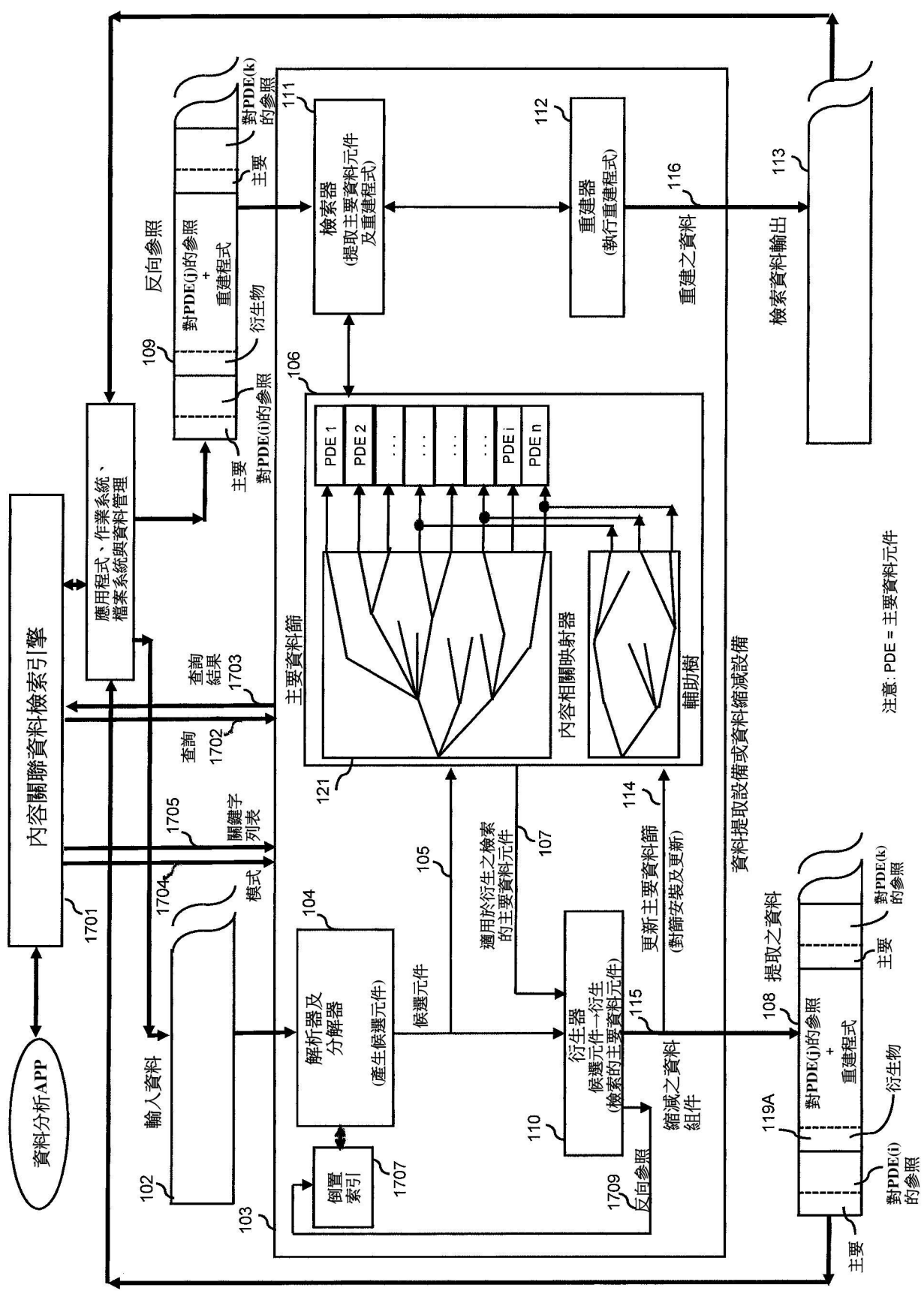
Prime, Reference to PDE PRINRACQNYCFEB ...
Prime, Reference to PDE NIKESHOELAHJUN ...
Derivative, Reference to PDE NIKESHOELAHJUN ..., Reconstitution Program 7
Prime, Reference to PDE WILSBALLAHNOV ...
.....
Prime, Reference to PDE NIKESHOELAHJUN ...
Prime, Reference to PDE NIKEJERSLAHOCT ...
Prime, Reference to PDE ADIDSHOESJCSEP ...



第 16E 圖

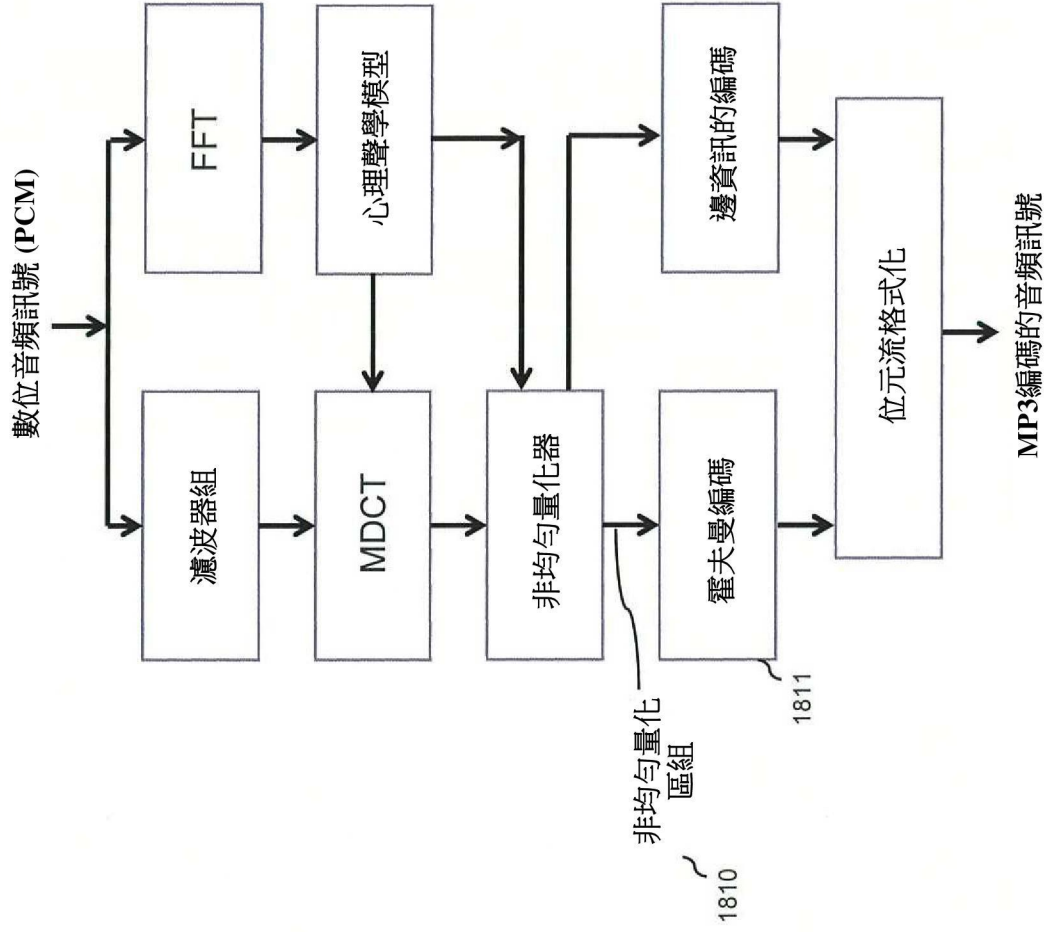


第17圖



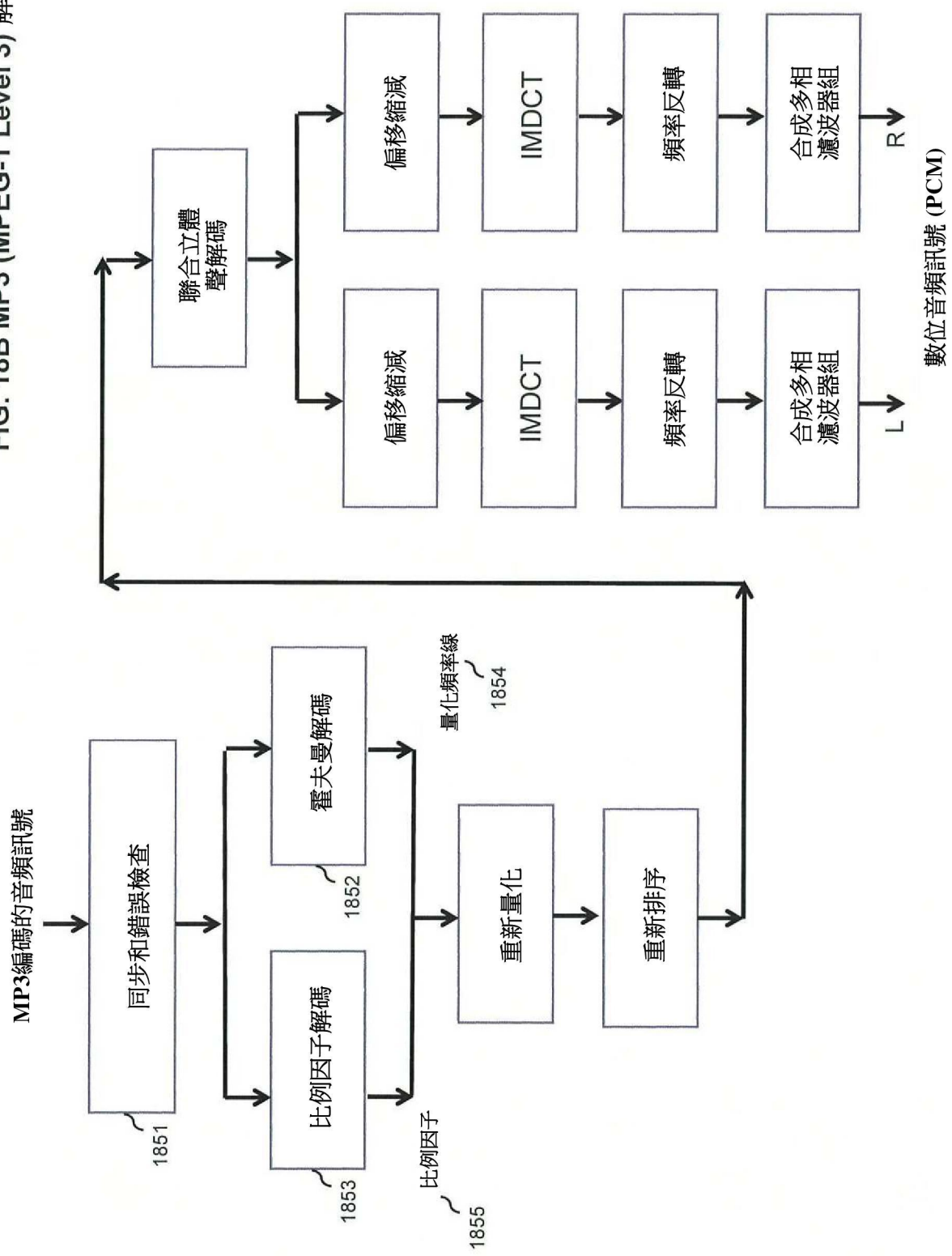
第 18A 圖

MP3 (MPEG-1 Level 3) 編碼器



第 18B 圖

FIG. 18B MP3 (MPEG-1 Level 3) 解碼器





第19圖

