



(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 I626547 B

(45)公告日：中華民國 107 (2018) 年 06 月 11 日

(21)申請案號：103107071

(22)申請日：中華民國 103 (2014) 年 03 月 03 日

(51)Int. Cl. : G06F17/30 (2006.01)

G06F11/16 (2006.01)

(71)申請人：國立清華大學(中華民國) NATIONAL TSING HUA UNIVERSITY (TW)
新竹市光復路 2 段 101 號(72)發明人：蕭宏章 HSIAO, HUNG CHANG (TW)；廖啓村 LIAO, CHI TSUN (TW)；蔡嘉平
TSAI, CHIA PING (TW)；鍾葉青 CHUNG, YEH CHING (TW)

(74)代理人：楊敏玲

(56)參考文獻：

TW 200743965A

TW 200823682A

US 8098511B2

US 2010/0161565A1

US 2010/0306486A1

審查人員：林琮烈

申請專利範圍項數：5 項 圖式數：4 共 20 頁

(54)名稱

於分散式資料庫中將系統狀態一致地還原至欲還原時間點之方法及系統

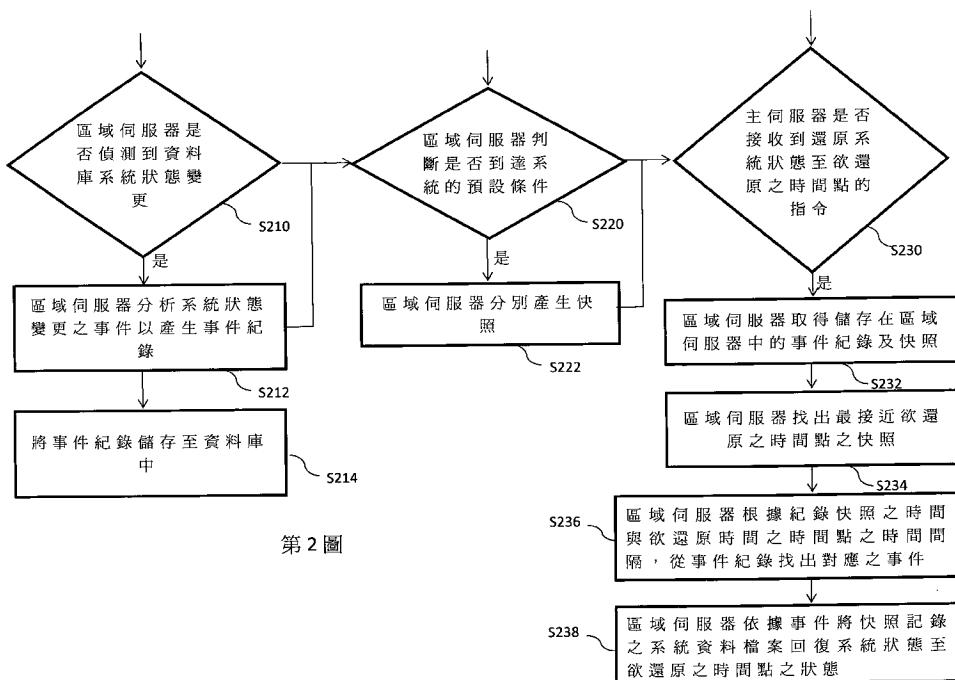
SYSTEM AND METHOD FOR RECOVERING SYSTEM STATE CONSISTENCY TO ANY POINT-IN-TIME IN DISTRIBUTED DATABASE

(57)摘要

一種於具有主伺服器及複數區域伺服器之分散式資料庫中，將主伺服器及區域伺服器之系統狀態一致地還原至欲還原之時間點之方法，包括當區域伺服器偵測到資料庫中系統狀態變更時，分析變更事件以產生事件紀錄並儲存至資料庫中。經過預設條件，區域伺服器產生快照。當主伺服器接收到還原系統狀態至欲還原之時間點的指令時，指示區域伺服器取得儲存在區域伺服器中的事件紀錄及快照、找出最接近欲還原之時間點之快照，並根據快照之時間與欲還原時間之時間點之間隔，從事件紀錄找出對應之事件及快照，以回復系統狀態至欲還原之時間點。

A method for recovering system state consistency to any point-in-time in distributed database, wherein the distributed database comprises a primary server and a plurality of region servers, comprising: analyzing the event changes to generate an event log when the region servers detect the system state changes; after default condition, the region servers generate a snapshot respectively; when the primary server receives the instruction to recover system state to designed point-in-time, indicates the region servers to implement: read the event log and snapshots stored in region servers; finding the snapshot closest to the designed point-in-time; finding the event log and snapshot correspond to the interval between the time recorded in snapshot and the designed point-in-time to recover system state to the designed point-in-time.

指定代表圖：



發明專利說明書

(本說明書格式、順序，請勿任意更動)

【發明名稱】(中文/英文)

於分散式資料庫中將系統狀態一致地還原至欲還原時間點之方法及系統/

System and method for recovering system state consistency to any point-in-time
in distributed database

【技術領域】

【0001】 一種還原系統狀態之方法及系統，特別是一種於分散式資料庫中回復系統狀態至同一時間點的方法及系統。

【先前技術】

【0002】 網路技術的快速演進改變了人類使用網路的習慣，且透過網路的交流，世界各地的人們可以即時、零距離的分享資訊。舉例而言，多數網站已轉變成Web2.0的型態，使用Web2.0與傳統網站最大的差異在於資料的來源，Web2.0的經營者通常提供的為資訊平台，並藉由人與人之間的分享達到資料傳播的目的。雖然Web2.0可以透過知識的累積取得更及時、充足的資料，相對地，由於使用者長時間提供資訊，累積的資料量遠大於傳統網站所需的資源，造成系統容量不堪負荷。

【0003】 為了解決上述的問題，非關聯式資料庫(No SQL)的技術逐漸應用於儲存各種海量資料上。相較於傳統的關聯式資料庫，非關聯式資料庫的優點是可水平擴張，亦即此種資料庫的容量彈性，當資料庫負荷過重時可隨時擴張調整其容量。常見的非關聯式資料庫為HBase，HBase為分散式資料庫，並運用Apache Hadoop作為檔案系統。於HBase中又分為主伺服器及區域伺服器，主伺服器與區域伺服器相互連結，當主伺服器偵測到其

中之一之區域伺服器毀損時，可還原此區域之資料至先前儲存之狀態。

【0004】 在現有的HBase系統中，雖然主伺服器自動偵測區域伺服器毀損並回復資料之先前狀態，但無法將區域伺服器的分佈等其他系統狀態回復至特定的時間點；此外，回復系統至特定時間點所依據的歷史紀錄係為一段時間內之紀錄，也就是說，當回復系統時，僅能針以一段時間作為單位進行回復，而無法回復特定時間點的狀態。基於上述內容，本技術領域亟需一種於非關聯資料庫中還原系統狀態至相同時間點之方法，並且保證系統狀態恢復能具有一致性。

【發明內容】

【0005】 發明內容旨在提供本揭示內容的簡化摘要，以使閱讀者對本揭示內容具備基本的理解。此發明內容並非本揭示內容的完整概述，且其用意並非在指出本發明實施例的重要/關鍵元件或界定本發明的範圍。

【0006】 本發明之一態樣係有關一種於具有主伺服器及複數個區域伺服器之分散式資料庫中，將主伺服器及區域伺服器之系統狀態一致地還原至欲還原之時間點之方法此方法包括步驟：當區域伺服器之一偵測到資料庫中每一系統狀態變更時，區域伺服器分析系統狀態變更之事件以產生事件紀錄，並將事件紀錄儲存至資料庫中，其中，事件紀錄分別具有時間向量用以判斷事件發生之順序；經過預設條件後，區域伺服器分別產生快照，其中快照係為格式化的系統資料檔案；以及當主伺服器接收到還原系統狀態至欲還原之時間點的指令時，指示區域伺服器執行下列步驟：取得儲存在區域伺服器中的事件紀錄及快照、找出最接近欲還原之時間點之快照、根據紀錄快照之時間與欲還原時間之時間點之時間間隔，從事件紀錄

找出對應之事件，並依據事件將快照紀錄之系統資料檔案回復系統狀態至欲還原之時間點之狀態。

【0007】 依據本發明之一實施例，其中，時間向量具有區域編碼及全域編碼。

【0008】 依據本發明之又一實施例，於此方法之步驟(3)，藉由時間向量找出對應欲還原之時間點之事件紀錄更包含下列步驟：判斷快照時間與欲還原之時間點之時間間隔、根據事件紀錄之全域編碼判斷事件於分散式資料庫中發生之時間，並根據事件紀錄之區域編碼判斷事件於對應全域編碼之時間內發生之一順序；以及依據快照及找出之順序及事件，將系統狀態回復至欲還原之時間點。

【0009】 依據本發明之另一實施例，步驟產生事件紀錄更包括步驟：區域伺服器接收到系統狀態變更指令、解析變更指令、計算區域/全域編碼並且產生事件紀錄；以及區域伺服器將事件紀錄儲存於資料庫中。

【0010】 依據本發明之又一實施例，快照更包含對應該些事件之該些檔案。

【0011】 一種於分散式資料庫中還原系統狀態至相同一時間點之系統，包括：主伺服器，具有主處理器、主記憶體以及主儲存裝置；以及複數區域伺服器，分別具有區域伺服器處理器、區域伺服器記憶體以及區域伺服器儲存裝置；其中，當主伺服器接收到還原系統狀態至欲還原之時間點的指令時，指示區域伺服器執行下列步驟：取得儲存在區域伺服器中的事件紀錄及快照、找出最接近欲還原之時間點之快照；以及根據紀錄快照之時間與欲還原時間之時間點之時間間隔，從該些事件紀錄找出對應之事

件，並依據事件將快照紀錄之系統資料檔案回復系統狀態至欲還原之時間點之狀態。

【0012】 在參閱下文實施方式後，本發明所屬技術領域中具有通常知識者當可輕易瞭解本發明之基本精神及其他發明目的，以及本發明所採用之技術手段與實施態樣。

【圖式簡單說明】

【0013】 第1圖為依照本發明系統一實施方式所繪示的系統架構圖；

【0014】 第2圖為依照本發明方法一實施方式所繪示的流程圖；

【0015】 第3圖為依照本發明方法另一實施方式所繪示的流程圖；以及

【0016】 第4圖為依照本發明方法又一實施方式所繪示的流程圖。

【實施方式】

【0017】 為了使本揭示內容的敘述更加詳盡與完備，下文針對了本發明的實施態樣與具體實施例提出了說明性的描述；但這並非實施或運用本發明具體實施例的唯一形式。實施方式中涵蓋了多個具體實施例的特徵以及用以建構與操作這些具體實施例的方法步驟與其順序。然而，亦可利用其他具體實施例來達成相同或均等的功能與步驟順序。

【0018】 除非本說明書另有定義，此處所用的科學與技術詞彙之含義與本發明所屬技術領域中具有通常知識者所理解與慣用的意義相同。此外，在不和上下文衝突的情形下，本說明書所用的單數名詞涵蓋該名詞的複數型；而所用的複數名詞時亦涵蓋該名詞的單數型。

【0019】 於本說明書內所述的「分散式資料庫」係為多台主機所結

合而成，可讓多使用者分享結構化的資料表。此外，於本實施例中之分散式資料庫為主從式架構，並將系統劃分為數個資料區塊。於系統內具有一主伺服器以及複數區域伺服器，主伺服器係用以管理區域伺服器的溝通、協調以及檔案的目錄，區域伺服器則各自負責一個資料區塊，管理資料區塊內資料的存取。

【0020】 於本說明書內所述的「系統狀態」係指系統內檔案之狀態和區域伺服器管理資料區塊之狀態，當資料表格經由使用者新增、刪除或修改，或者當區域伺服器進行資料整理搬移或經由使用者重設資料範圍、關閉或新增等資料異動後，檔案狀態、各個資料區塊、區域伺服器狀態即已不同，此時，系統狀態已改變。舉例而言，檔案在系統內之編排、放置位置皆由區域伺服器執行處理，當檔案僅是搬移至不同的區域伺服器中，但檔案未經由使用者新增、修改或刪除之情形下，雖然對使用者而言，所見的檔案狀態是相同的，但是區域伺服器管理的資料範圍會因此變更，因此區域伺服器之狀態與檔案搬移前不同，系統狀態因而改變。

【0021】 每一次檔案新增、檔案刪除、檔案修改或檔案搬移等，皆視為本說明書內所述的「事件」。而「事件紀錄」即為記錄這些事件發生的狀況，舉例但不限於因事件造成改變的檔案識別名稱或位址、事件發生的內容、事件發生的時間。其中，複數筆事件紀錄亦可封裝成單一檔案，並供系統存取。

【0022】 於本說明書內所述的事件紀錄更包括「區域編碼」及「全域編碼」，區域編碼係依據區域伺服器處理的事件順序給予編碼，舉例而言，於本實施例中所使用的編碼為依序由小到大的流水號，根據此流水號之大

10具有一主處理器12、一主記憶體16以及一主儲存裝置18。主處理器12舉例但不限於中央處理器，於本實施例中，主處理器12具有兩個主控制器14a、14b，控制器14a、14b係用以處理資料的流進與流出，當其中一主控制器14a失效時，主控制器14b會偵測到主控制器14a停止運作，並接管主控制器14a的工作。主記憶體16係用來暫存更新的目錄，當經過預設的時間時（例如：預設的時間區間或預設條件，舉例而言，每小時或每發生一百次事件後，將主記憶體16內之資料寫入檔案系統18中），主控制器14a或主控制器14b會把主記憶體16裡的目錄存入至檔案系統18。

【0027】 區域伺服器20分別負責一資料區塊內資料的新增、刪除、修改及移動。以區域伺服器20為例，具有一區域處理器22、一事件記憶體26此區域伺服器20並可讀取檔案系統18。於本實施例中，區域處理器22具有多個區域記憶體（於第1圖中僅顯示兩個區域記憶體24a、24b）。區域處理器22係用以處理資料的流進與流出。區域記憶體24a、24b 係用來暫存更新操作，一定量或一段時間後會將區域記憶體內的資料寫入檔案系統18中。事件記憶體26係用以處理事件紀錄，當經過預設的時間時（例如：預設的時間區間或預設條件，舉例而言，每小時或每發生一百次事件後，將事件記憶體26內之資料寫入檔案系統18中）。

【0028】 主伺服器10雖用來管理系統中資料的目錄及協調區域伺服器20間的分工合作，但於其他實施例中，主伺服器10亦可同時兼具有區域伺服器20管理資料之功能。

【0029】 又如第1圖中，當使用者1欲讀取表格資料時，使用者1連至區域伺服器20，區域伺服器20之區域處理器22從區域記憶體24a、24b和檔

案系統18搜尋表格資料，區域處理器22並回傳此表格資料給使用者1。

【0030】 當使用者1欲寫入資料至表格時，區域伺服器20收到指令後，會由區域處理器22傳遞檔案至區域記憶體24a、24b中，並產生事件紀錄至事件記憶體26，之後，區域處理器22會回傳傳送成功的訊息至使用者1。

【0031】 第2圖為依照本發明方法一實施方式所繪示的流程圖，本實施例之流程係執行於第1圖之系統架構上。本方法包括步驟區域伺服器是否偵測到資料庫系統狀態變更S210、區域伺服器分析系統狀態變更之事件以產生事件紀錄S212、將事件紀錄儲存至資料庫中S214、區域伺服器判斷是否到達系統的預設條件S220、區域伺服器分別產生快照S222、主伺服器是否接收還原系統狀態至欲還原之時間點的指令S230、區域伺服器取得儲存在區域伺服器中的事件紀錄及快照S232、區域伺服器找出最接近欲還原之時間點之快照，S234、區域伺服器根據紀錄快照之時間與欲還原時間之時間點之時間間隔，從事件紀錄找出對應之事件S236。以及區域伺服器依據事件將快照記錄之系統資料檔案回復系統狀態至欲還原之時間點之狀態S238。

【0032】 步驟S210係為區域伺服器判斷是否有變更系統狀態的指令，當系統接收到會產生系統狀態變更的指令時，區域伺服器會分析此事件，並依據系統變更的性質（如：檔案之新增、刪除、搬移）、變更的對象（即，哪個檔案被變更）、執行的時間與操作的順序等資訊產生事件紀錄，並將此事件紀錄儲存至資料庫中S214。

【0033】 又，系統步驟S220為判斷區域伺服器判斷是否到達系統的預設條件，若符合系統之預設條件，區域伺服器會產生快照S222。以記錄

當時之系統狀態，當資料庫需進行系統還原時，得以據此回復檔案資料。

【0034】 此外，步驟S230為判斷主伺服器是否接收到還原系統狀態至欲還原之時間點的指令，當該主伺服器偵測到系統還原指令時，主伺服器將指示區域伺服器取得儲存在區域伺服器中的事件紀錄及快照S232、區域伺服器找出最接近欲還原之時間點之快照S234、區域伺服器根據紀錄快照之時間與欲還原時間之時間點之時間間隔，從事件紀錄找出對應之事件S236以及區域伺服器依據事件將快照記錄之系統資料檔案回復系統狀態至欲還原之時間點之狀態S238。其中，於其他實施例中，步驟S210、S220及步驟S230之順序可以相互交換，並不限於此。

【0035】 第3圖為依照本發明方法另一實施方式所繪示的流程圖，如圖所示，於步驟S236包含步驟判斷快照時間與欲還原之時間點之時間間隔S2362、根據事件紀錄之全域編碼判斷事件於分散式資料庫中發生之時間，並根據事件紀錄之區域編碼判斷事件對應全域編碼之時間內發生之順序S2364、依據快照及找出之順序及事件，將系統狀態回復至欲還原之時間點S2366。

【0036】 第4圖為依照本發明方法又一實施方式所繪示的流程圖。其中，步驟212包含步驟：區域伺服器接收到系統狀態變更指令S2122以及解析變更指令、計算區域/全域編碼並且產生事件紀錄S2124。

【0037】 本揭示內容旨在揭示一種於分散式資料庫中還原系統狀態至相同時間點之方法及系統，此方法與系統可使資料庫回復至相同時間點，降低資料錯誤、遺失的風險。

【0038】 雖然上文實施方式中揭露了本發明的具體實施例，然其並非

用以限定本發明，本發明所屬技術領域中具有通常知識者，在不悖離本發明之原理與精神的情形下，當可對其進行各種更動與修飾，因此本發明之保護範圍當以附隨申請專利範圍所界定者為準。

【符號說明】

【0001】	1	使用者
【0002】	10	主伺服器
【0003】	12	主處理器
【0004】	14a、14b	主控制器
【0005】	16	主記憶體
【0006】	18	檔案系統
【0007】	20	區域伺服器
【0008】	22	區域處理器
【0009】	24a、24b	區域記憶體
【0010】	26	事件記憶體

發明摘要

※ 申請案號：103107071

※ 申請日：103/03/03

※IPC 分類：
G06F 17/30 (2006.01)
G06F 11/16 (2006.01)

【發明名稱】(中文/英文)

於分散式資料庫中將系統狀態一致地還原至欲還原時間點之方法及系統 / System and method for recovering system state consistency to any point-in-time in distributed database

【中文】

一種於具有主伺服器及複數區域伺服器之分散式資料庫中，將主伺服器及區域伺服器之系統狀態一致地還原至欲還原之時間點之方法，包括當區域伺服器偵測到資料庫中系統狀態變更時，分析變更事件以產生事件紀錄並儲存至資料庫中。經過預設條件，區域伺服器產生快照。當主伺服器接收到還原系統狀態至欲還原之時間點的指令時，指示區域伺服器取得儲存在區域伺服器中的事件紀錄及快照、找出最接近欲還原之時間點之快照，並根據快照之時間與欲還原時間之時間點之時間間隔，從事件紀錄找出對應之事件及快照，以回復系統狀態至欲還原之時間點。

【英文】

A method for recovering system state consistency to any point-in-time in distributed database, wherein the distributed database comprises a primary server and a plurality of region servers, comprising: analyzing the event changes to generate an event log when the region servers detect the system state changes;

after default condition, the region servers generate a snapshot respectively; when the primary server receives the instruction to recover system state to designed point-in-time, indicates the region servers to implement: read the event log and snapshots stored in region servers; finding the snapshot closest to the designed point-in-time; finding the event log and snapshot correspond to the interval between the time recorded in snapshot and the designed point-in-time to recover system state to the designed point-in-time.

【代表圖】

【本案指定代表圖】：第（ 2 ）圖。

【本代表圖之符號簡單說明】：

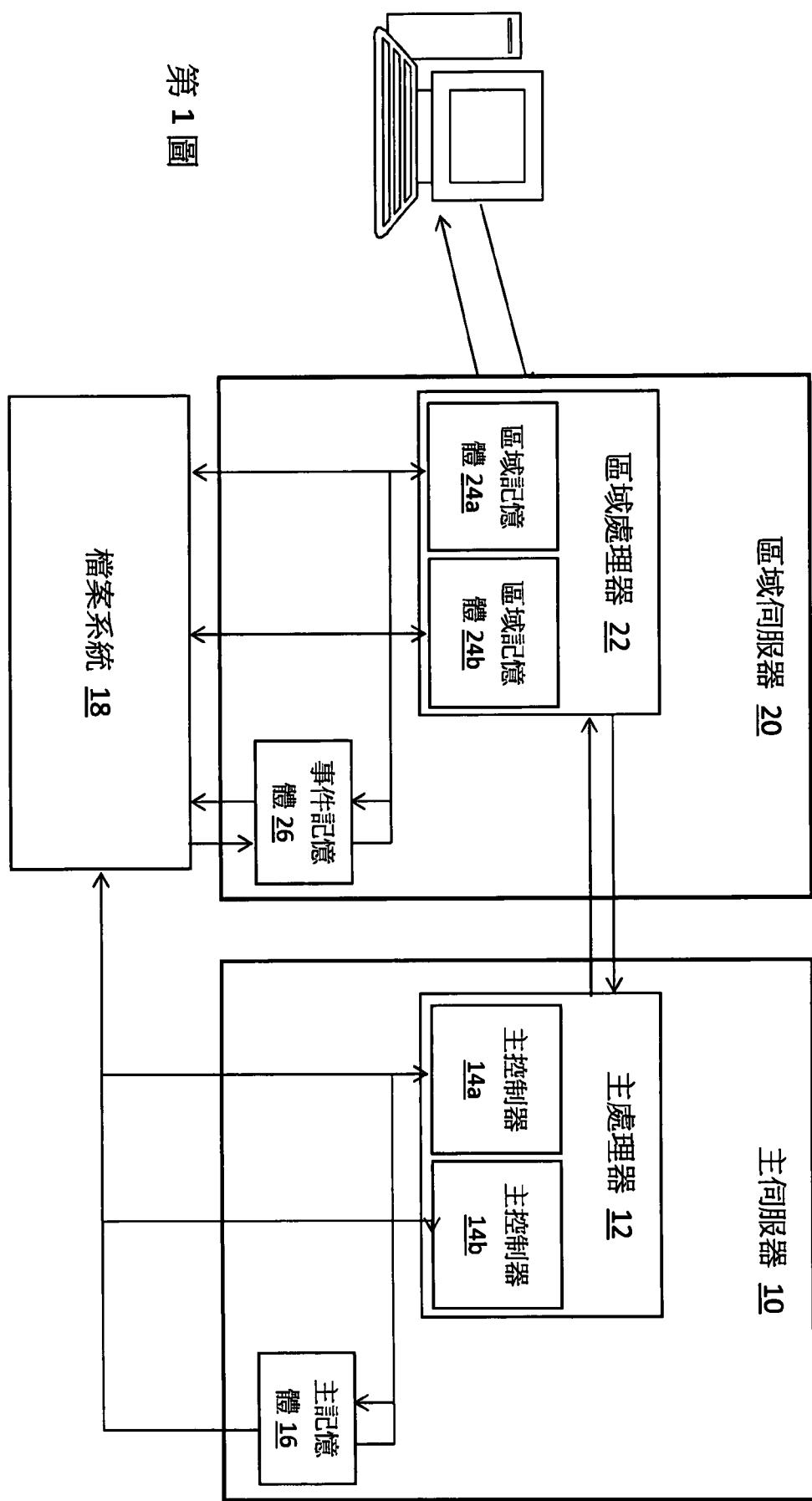
【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

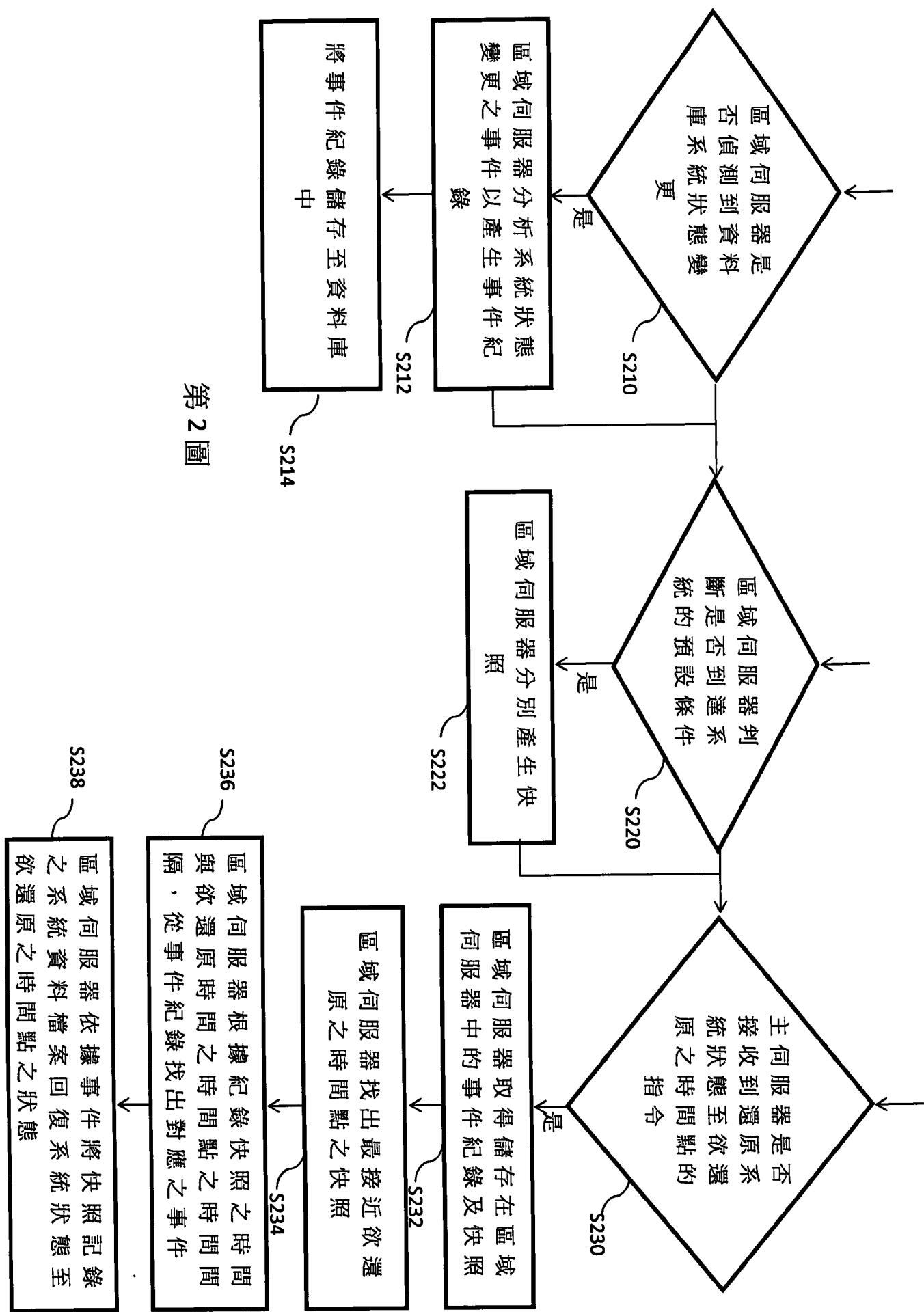
I626547

圖式

如附件所示

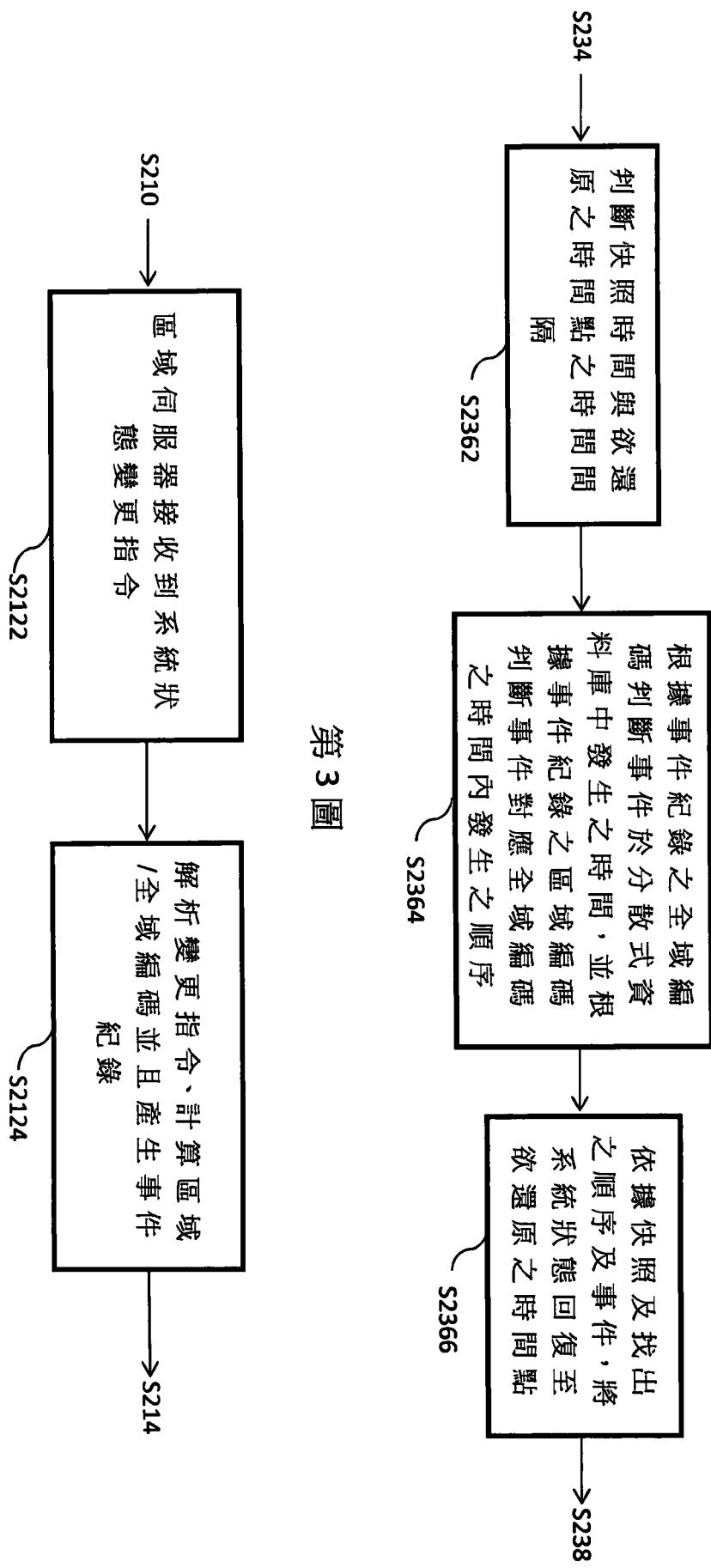
第 1 圖





第 2 圖

I626547



after default condition, the region servers generate a snapshot respectively; when the primary server receives the instruction to recover system state to designed point-in-time, indicates the region servers to implement: read the event log and snapshots stored in region servers; finding the snapshot closest to the designed point-in-time; finding the event log and snapshot correspond to the interval between the time recorded in snapshot and the designed point-in-time to recover system state to the designed point-in-time.

【代表圖】

【本案指定代表圖】：第（ 2 ）圖。

【本代表圖之符號簡單說明】：

【本案若有化學式時，請揭示最能顯示發明特徵的化學式】：

無劃線修正頁

小可以判定事件發生之先後順序。全域編碼則為依據整體系統的時間給予編碼。對於整體系統而言，全域編碼亦可作為改變區域伺服器狀態之事件順序的編碼；區域編碼為使用者操作表格之事件編碼。

【0023】 於本說明書內所述的「預設條件」係依據不同系統而有所差異，舉例而言，可以依照預設的時間區間（例如但不限於：每小時產生快照一次），或者預設條件（例如但不限於：每發生一百次事件），但不限於此。

【0024】 於本說明書內所述的「快照」係為格式化的系統資料檔案，用以記錄系統在某一時刻的狀態。此外，為避免冗長的記錄過程及降低占用系統空間，每次快照會間隔一長時間。假設系統每天產生快照一次，此次產生的快照即為此時此刻的系統狀態。換言之，日後可根據此快照將系統還原至產生快照之時間點。於一實施例中，系統產生之快照包含預設條件內之檔案，亦即，包括新增、修改或刪除完畢的檔案。另外也包含區域伺服器的狀態。

【0025】 於本說明書內所述的「目錄」係用以記錄檔案存放於檔案系統中的實體位置，區域伺服器可透過讀取目錄取得檔案的實體位置，並進一步進行存取。

【0026】 請參考第1圖，第1圖為依照本發明系統一實施方式所繪示的系統架構圖。於本系統中，具有一主伺服器10、複數個區域伺服器20（於第1圖中僅顯示一個區域伺服器），其中，主伺服器10係用以管理整個系統狀態，此外，當偵測到區域伺服器20停止運作時，主伺服器10會施放指令，協調其他運作中的區域伺服器接管停止的區域伺服器20。其中，主伺服器

申請專利範圍

1. 一種於具有一主伺服器及複數個區域伺服器之分散式資料庫中，將該主伺服器及該些區域伺服器之系統狀態一致地還原至欲還原之一時間點之方法，包括：

當該些區域伺服器之一偵測到該分散式資料庫中每一系統狀態變更時，該些區域伺服器分析系統狀態變更之一事件以產生一事件紀錄，並且該些區域伺服器將該事件紀錄儲存至該分散式資料庫中，其中，該些事件紀錄分別具有一時間向量用以判斷該些事件發生之順序；經過預設條件後，該些區域伺服器分別產生一快照，其中該快照係為格式化的系統資料檔案；以及

當該主伺服器接收到還原系統狀態至欲還原之一時間點的指令時，該主伺服器指示該些區域伺服器執行下列步驟：

- (1)取得儲存在該些區域伺服器中的該些事件紀錄及該些快照；
- (2)找出最接近欲還原之該時間點之該快照；以及
- (3)根據記錄該快照之時間與欲還原時間之該時間點之時間間隔，從該些事件紀錄找出對應之該些事件，並依據該些事件將該快照記錄之系統資料檔案回復系統狀態至欲還原之該時間點之狀態；

其中，該些時間向量具有一全域編碼及一區域編碼，該全域編碼為依據整體系統的時間給予編碼，該區域編碼係依據區域伺服器處理的事件順序給予編碼。

2. 如請求項1所述之方法，其中，步驟(3)包含下列步驟：

區域伺服器判斷該快照時間與該欲還原之該時間點之時間間隔；

區域伺服器根據該些事件紀錄之全域編碼判斷該些事件於該分散式資料庫中發生之時間，並根據該些事件紀錄之區域編碼判斷該些事件於該對應全域編碼之時間內發生之一順序；以及
區域伺服器依據該快照及找出之該順序及該些事件，將系統狀態回復至欲還原之該時間點。

3. 如請求項1所述之方法，步驟產生該事件紀錄更包括下列步驟：

區域伺服器接收到一系統狀態變更指令；以及
區域伺服器解析該變更指令、計算區域/全域編碼並且產生該事件紀錄。

4. 如請求項1所述之方法，該快照更包含對應該些事件之該些系統資料檔案。

5. 一種於一分散式資料庫中還原系統狀態至相同一時間點之系統，包括：
一主伺服器，具有至少一主處理器、至少一主記憶體以及至少一主儲存裝置；以及
複數區域伺服器，分別具有至少一區域伺服器處理器、至少一區域伺服器記憶體以及至少一區域伺服器儲存裝置；

其中，當該些區域伺服器之一偵測到該分散式資料庫中每一系統狀態變更時，該些區域伺服器分析系統狀態變更之一事件以產生一事件紀錄，並並且該些區域伺服器將該事件紀錄儲存至該分散式資料庫中，其中，該些事件紀錄分別具有一時間向量用以判斷該些事件發生之順序；
其中，當該主伺服器接收到還原系統狀態至欲還原之一時間點的指令時，該主伺服器指示該些區域伺服器執行下列步驟：

- (1)取得儲存在該些區域伺服器中的該些事件紀錄及該些快照；

106年12月12日修正替換頁

(2)找出最接近欲還原之該時間點之該快照；以及
(3)根據紀錄該快照之時間與欲還原時間之該時間點之時間間隔，從該些事件紀錄找出對應之該些事件，並依據該些事件將該快照紀錄之系統資料檔案回復系統狀態至欲還原之該時間點之狀態；其中，該些時間向量具有一全域編碼及一區域編碼，該全域編碼為依據整體系統的時間給予編碼，該區域編碼係依據區域伺服器處理的事件順序給予編碼。