(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 November 2004 (11.11.2004)

PCT

(10) International Publication Number
WO 2004/097677 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number:
PCT/IB2004/000669

(22) International Filing Date: 4 March 2004 (04.03.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
03405295.1 28 April 2003 (28.04.2003) EP

(71) Applicant *(for all designated States except US)*: **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, NY 10504 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **HILD, Stefan, G.** [DE/US]; 130 Mitchell Road, Somers, NY 10589 (US). **PAWLITZEK, René, A.** [LI/CH]; Schlossbergstrasse 17, CH-8802 Kilchberg (CH). **RJAIBI, Walid** [CA/CA]; 10 Monkhouse Road, Markham, Ontario L6E 1E9 (CA). **STOLZE, Markus** [DE/CH]; Sustenweg 40, CH-3014 Bern (CH).

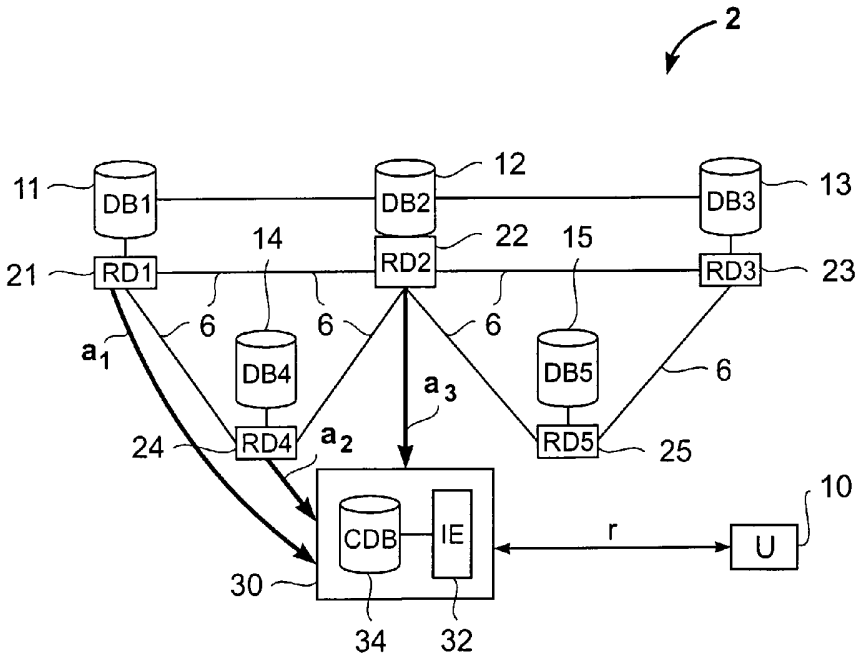
(74) Agents: **KLETT, Peter, M.** et al.; IBM Research GmbH, Zurich Research Laboratory, Intellectual Property Law, Saeumerstrasse 4 / Postfach, CH-8803 Rueschlikon (CH).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),

[Continued on next page]

(54) Title: AUTOMATIC DATA CONSOLIDATION



(57) Abstract: The present invention discloses a method, request detector, inference engine, and system for consolidating data from distributed databases into a central database. The method comprises the steps of receiving access information comprising request information to the distributed databases, analyzing the received access information, and aggregating into the central database the data content of the distributed databases in dependence on the analyzed access information.

WO 2004/097677 A1

Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

AUTOMATIC DATA CONSOLIDATION

TECHNICAL FIELD

The present invention is related to a method, apparatus, and system for consolidating data from distributed databases into a central database.

BACKGROUND OF THE INVENTION

Today many organizations concentrate their IT (information technology) spending on methods and technologies that help them reduce cost by increasing the efficiency and effectiveness of their IT infrastructure. A key pain point faced by many organization and companies is that, in the beginning of the Internet boom, also called the "dot-com" bubble, many organizations had to embrace the Internet overnight, and in doing so established infrastructure elements in an ad-hoc fashion that were not well-designed for scalability or growth. Today, these organizations and companies are faced with infrastructures that are loosely assembled, are costly to maintain, and are difficult and expensive to grow with the business needs. This is evident in both the business processes as well as in the way these organizations manage data.

In many instances, the data is fragmented within the organization or company, with different database systems being utilized in different departments, all of which often maintain essentially the same data in multiple formats using different database table designs. Here, it would be hugely beneficial to maintain all data in a (logically) single place using a standardized schema. Having such centralized data warehouse or database engine would enable quick data analysis for improved customer relationship management, simplify development of new products, and reduce maintenance cost for the IT infrastructure itself while improving the reliability and availability of the entire system.

The international publication WO 99/52047A1 relates to a method and system for migrating data from one or more ASCII files and/or from one or more relational databases to one or more relational database tables without the need to write code. This allows the user to define

mapping templates and conditionals to assist in translating and transforming data values. The method also enforces referential integrity, data dependencies, order of operations, and uniqueness constraints using a predefined set of migration rules templates that are based on the principles of relational design. The method uses these mapping and migration rules templates to generate instructions for updating or populating relational database destination tables. The instructions control the data transfer, data translation, data transformation, data validation, foreign key insertion, and the addition of required codes and flags in the destination tables. A migration engine of the system includes a data map architect and an update processor which spawns the templates and migrates the data dynamically, utilizing the data definitions for the destination tables.

This prior art has the drawback that it is limited to a particular type of database systems, in which Oracle® Application tables are implemented. This requires users to manually define rules for the migration or at least interactively, i.e. by user interaction. (Oracle is a trademark of Oracle Corporation). This prior art does not show provisions which allow to discover databases across an organization or augment a consolidation process with actual access patterns.

From the above follows that there is still a need in the art for an efficient scheme that allows to consolidate data from distributed databases into a central database.

SUMMARY AND ADVANTAGES OF THE INVENTION

Disclosed is a scheme that performs two basic tasks, in which a) it monitors existing database access patterns in order to derive an overall view of the available data sources within an organization and how they are used, and b) over time it aggregates the data content of the various data sources into a new centralized repository and redirects calls to the remote database servers to this central database.

For that, three infrastructure elements are employed. So-called sensors, also referred to as interceptors or request detectors, monitor any data access across an organization or network. An inference engine that analyses the access patterns and the data formats contained in the

individual databases. A central database, also referred to as central data warehouse or centralized repository, aggregates the data from the individual data sources with the view to replace those eventually.

In the following the single infrastructure elements are explained in more detail.

The sensors or request detector are attached to infrastructure elements, typically software drivers that manage the database access of users and/or applications, and record the requests that are being submitted by users and applications. A typical example of such sensor is a modified ODBC (JDBC) driver. For instance, JDBC drivers are Java code that are frequently used today to access databases from Java programs. By modifying the JDBC driver the sensor logic can record all data requests that are initiated from programs or users to databases, and also which database is being addressed. Other examples can be derived by modifying the database itself. All data access is logged and transmitted to the inference engine, either in real time or in batch mode.

The inference engine analysis the data access recorded by the sensors or request detectors to identify i) the database engines used which are distributed databases within a network or networks and ii) the data schemes employed; for example, the inference engine learns what the format of the data tables is in the various database engines, what primary keys and foreign keys are employed, and what type of data is contained within those databases tables. Further, the inference engine can perform a correlation iii) between different databases; for example, the inference engine should correlate columns from different databases even though they may not be named the same.

Based on the results of the inference the engine generates a new data schema, generates an instance of that schema on the central data warehouse, i.e. the central database. Over time the inference engine then copies existing data from the individual distributed databases that have been discovered into this new central database. When completed, the inference engine may issue an order to redirect calls to the individual databases to the central data warehouse. This can be done by advising the request detectors to intercept the individual data access calls and redirecting them to the central database.

The central data warehouse or cental database is a database engine, e.g. an IBM DB/2. For increased availability a cluster may be utilized.

In accordance with the present invention, there is provided a method for consolidating data automatically from distributed databases into a central database. The method comprises the steps of receiving access information comprising request information to the distributed databases, analyzing the received access information, and aggregating into the central database the data content of the distributed databases in dependence on the analyzed access information. This allows a simple automatic migration of redundant data distributed over several databases.

The method can further comprise the steps of filtering the request information to the respective distributed databases from data traffic and forwarding the filtered request information within the access information to an inference engine. All the collected request information can be analyzed in one place, i.e. the information from the various databases can be compared and possible consolidations can be investigated.

For the central database a new data schema based on the analyzed access information can be generated. This has the advantage that a consolidated scheme can be used that meets the needs of the various distributed databases.

The analyzing step can comprise the usage of log-file information. This is simple to perform and does not require any change in the infrastructure, but may not yield access data at the same level of detail as a sensor or request detector would detect.

In accordance with another aspect of the present invention, there is provided a request detector for supporting data consolidation from distributed databases into a central database. The request detector can comprise a detecting means for detecting request information to the distributed databases, a transforming means that derives access information from the detected request information, and a providing means that sends the access information to an inference engine.

The request detector can be provided at each of the distributed databases to be consolidated, preferably in form of a modified ODBC (JDBC) driver. The request detector may even be integrated into each of the distributed databases to be consolidated.

The request detector can comprise redirecting means for redirecting a request to an individual database to the central database. This has the advantage that the request is forwarded directly to the consolidated central database and the user may get more information than it is provided by the individual database.

In accordance with yet another aspect of the present invention, there is provided an inference engine for controlling data consolidation from distributed databases into a central database. The inference engine can comprise means for analyzing of access information that is received from distributed databases and comprises request information to the respective distributed databases.

The inference engine can comprise a correlation means for correlating columns and/or rows between different distributed databases, but also fields, records, and/or data structures can be correlated. This leads to a new schema that then can be used by the consolidated central database. The inference engine allows a simple migration of the data. Equivalent information or data is brought together and stored on one place. This helps to avoid doubles in distributed systems.

In accordance with a further aspect of the present invention, there is provided a system for consolidating data from distributed databases into a central database. The system comprises a request detector at each of the distributed databases to be consolidated for providing access information comprising request information to the distributed databases, an inference engine for analyzing the received access information, and a central database into which the data content of the distributed databases is aggregated in dependence on the analyzed access information.

DESCRIPTION OF THE DRAWINGS

Preferred embodiments of the invention are described in detail below, by way of example only, with reference to the following schematic drawings.

FIG. 1 shows a schematic illustration of a distributed database structure.

FIG. 2 shows a schematic illustration of database structure according to the present invention.

FIG. 3 a shows a schematic illustration of a request and access information flow.

FIG. 3 b shows a schematic illustration of a redirect flow.

FIG. 4 shows schematic illustration of a consolidation of two databases into a central database.

The drawings are provided for illustrative purpose only and do not necessarily represent practical examples of the present invention to scale.

DESCRIPTION OF EMBODIMENTS

Fig. 1 shows a schematic illustration of a distributed database structure 1 with distributed databases 11, 12, 13, 14, 15. The databases, also labeled with DBx, are connected and accessible via a network 6. A user 10, also labeled with U, accesses here three distributed databases 11, 12, 14 in order to receive information that is distributed and provided by a first database 11, a second database 12, and a third database 14. In detail, the user 10, i.e. a user's computer, sends a first request r_1 to the first database 11, a second request r_2 to the second database 12, and third request r_3 to the third database 14. The user 10 receives the respective responses from the distributed databases 11, 12, 14 which then can be evaluated. However, in the example, three requests r_1, r_2, r_3 are sent to get the desired information. Further, the maintenance of all the distributed databases 10, 11, 12, 13, 14, 15 with overlapping content is not efficient and effective.

The same reference signs are used within the description to denote the same or like parts.

Turning to Fig. 2, which shows a schematic illustration of a modified database structure 2 according to the present invention. The database structure further comprises request detectors 21, 22, 23, 24, 25 attached to each of the distributed databases 11, 12, 13, 14, 15, respectively. Further, there is provided and connected to the network 6 a central data unit 30. This unit comprises an inference engine 32 and a central database 34 which are connected to each other and to the network 6, and thus to the distributed databases 11, 12, 13, 14, 15.

It is assumed that the user 10 sends the same requests r_1, r_2, r_3 to the respective distributed databases 11, 12, 14 as described with reference to Fig. 1. Each of the requests r_1, r_2, r_3 to the respective distributed databases 11, 12, 14 is now detected by the respective request detectors 21, 22, 24. For that, the detectors 21, 22, 23, 24, 25 comprise of detecting means (not shown) for detecting such requests r_1, r_2, r_3 to the distributed databases 11, 12, 14. From the requests r_1, r_2, r_3, also referred to as request information, access information a_1, a_2, a_3 is derived indicating, e.g. a database address, inquiry details, etc.. The access information a_1, a_2, a_3 is then sent to the inference engine 32 as indicated in Fig. 2.

The central data unit 30, i.e. the inference engine 32, receives the access information a_1, a_2, a_3 comprising request information r_1, r_2, r_3 and analyzes the received access information a_1, a_2, a_3 by using correlation means for correlating columns and/or rows between different distributed databases 11, 12, 13, 14, 15. In dependence on the analyzed access information, the data content of the distributed databases 11, 12, 13, 14, 15 is aggregated into the central database 34.

After some time, the distributed databases 11, 12, 13, 14, 15 can be removed, since the content is then consolidated and stored in the cental database 34. This allows more redundancy in storage and better archiving possibilities.

For the user 10 and the system it is advantageous that after the consolidation only one request r is to be sent to the central data unit 30. In response to that request r a more complete data set can be provided to the user 10.

Fig. 3a shows a schematic illustration of a flow of a request r and access information a for current distributed database structures in order to establish the central database 34. In the example, the request r is sent by the user 10 to one distributed database 1x. The request detector 2x attached to the distributed database 1x receives this request r, transforms it to access information a and sends this access information a for purposes of analysis to the inference engine 32.

Fig. 3b shows a schematic illustration of a redirect flow. This is performed when the data consolidation has been performed successfully. Then, the request detector 2x is informed by the inference engine 32 about redirecting and any request r sent by the user 10 is redirected as a redirect request RR by the request detector 2x to the central database 34. The request inquiry is then answered by the consolidated central database 34.

Fig. 4 shows a schematic illustration of a consolidation of two distributed databases 11, 12 into a central database 34. The figure shows on the left hand side the content of the first database 11 and the content of the second database 12. The first database 11 stores "Names", like Joe, Bob and Alice, and their respective "Age", 28, 40, and 18. The second database 12 stores also "Names", here Joe and Bob, but is stores further the "Place of birth" instead of the Ages. As can be seen from this simple example, there are some overlapping names, i.e. information, like it is the case in many distributed database structures today. A data consolidation of the data is therefore desired saving infrastructure and maintenance costs. The inference engine 32 analyzes the available data, compares it, and performs a correlation. This correlation correlates columns and/or rows between different distributed databases to find similarities or matches. As indicated in the figure, the data fields "Name" appear in both distributed databases 11, 12. Thus, the distributed databases 11, 12 are candidates to merge their content into the central database 34. The inference engine 32 generates the new data schema and provides it to the central database 34. Finally, the central database 34 is filled with the content of the distributed databases 11, 12 to have the fields "Name", "Age", and "Place of birth" with the records for 'Joe, 28, Bern'; 'Bob 40, Zurich'; and 'Alice, 18'.

Computer program means or computer program in the present context mean any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after either

or both of the following a) conversion to another language, code or notation; b) reproduction in a different material form.

Any disclosed embodiment may be combined with one or several of the other embodiments shown and/or described. This is also possible for one or more features of the embodiments.

CLAIMS

1. A method for consolidating data from distributed databases (11, 12, 13, 14, 15) into a central database (34) comprising the steps of:

   - receiving access information (a_1, a_2, a_3) comprising request information (r_1, r_2, r_3) to the distributed databases (11, 12, 14),

   - analyzing the received access information (a_1, a_2, a_3), and
   - aggregating into the central database (34) the data content of the distributed databases (11, 12, 13, 14, 15) in dependence on the analyzed access information (a_1, a_2, a_3).

2. The method according to claim 1 further comprising providing a request detector (21, 22, 23, 24, 25) at each of the distributed databases (11, 12, 13, 14, 15) to be consolidated.

3. The method according to claim 2 further comprising the steps of filtering the request information (r_1, r_2, r_3) to the respective distributed databases (11, 12, 14) from data traffic and forwarding the filtered request information (r_1, r_2, r_3) within the access information (a_1, a_2, a_3) to an inference engine (32).

4. The method according to claim 2 further comprising the step of integrating the request detector (22) into each of the distributed databases (12) to be consolidated.

5. The method according to any preceding claim further comprising generating for the central database (34) a new data schema based on the analyzed access information (a_1, a_2, a_3).

6. The method according to claim 1, wherein the step of analyzing comprises using log-file information.

7. A computer program element comprising program code means for performing a method of any one of the claims 1 to 6 when said program is run on a computer.

8. A computer program product stored on a computer usable medium, comprising computer readable program means for causing a computer to perform a method according to anyone of the preceding claims 1 to 6.

9. A request detector (21, 22, 23, 24, 25) for supporting data consolidation from distributed databases (11, 12, 13, 14, 15) into a central database (34) comprising a detecting means for detecting request information (r_1, r_2, r_3) to the distributed databases (11, 12, 14), a transforming means that derives access information (a_1, a_2, a_3) from the detected request information (r_1, r_2, r_3), and a providing means that sends the access information (a_1, a_2, a_3) to an inference engine (32).

10. The request detector (2x) according to claim 9 further comprising redirecting means for redirecting a request (r) to an individual database (1x) to the central database (34).

11. An inference engine (32) for controlling data consolidation from distributed databases (11, 12, 13, 14, 15) into a central database (34) comprising means for analyzing of access information (a_1, a_2, a_3) that is received from distributed databases (11, 12, 14) and comprises request information (r_1, r_2, r_3) to the respective distributed databases (11, 12, 14).

12. The inference engine according to claim 11 further comprising correlation means for correlating columns and/or rows between different distributed databases (11, 12, 13, 14, 15).

13. A system for consolidating data from distributed databases (11, 12, 13, 14, 15;1x) into a central database (34) comprising:

- a request detector (21, 22, 23, 24, 25; 2x) at each of the distributed databases (11, 12, 13, 14, 15; 1x) to be consolidated for providing access information (a_1, a_2, a_3; a) comprising request information (r_1, r_2, r_3; r) to the distributed databases (11, 12, 14; 1x);

- an inference engine (32) for analyzing the received access information (a_1, a_2, a_3; a); and

- a central database (34) into which the data content of the distributed databases (11, 12, 13, 14, 15; 1x) is aggregated in dependence on the analyzed access information (a_1, a_2, a_3; a).

* * *

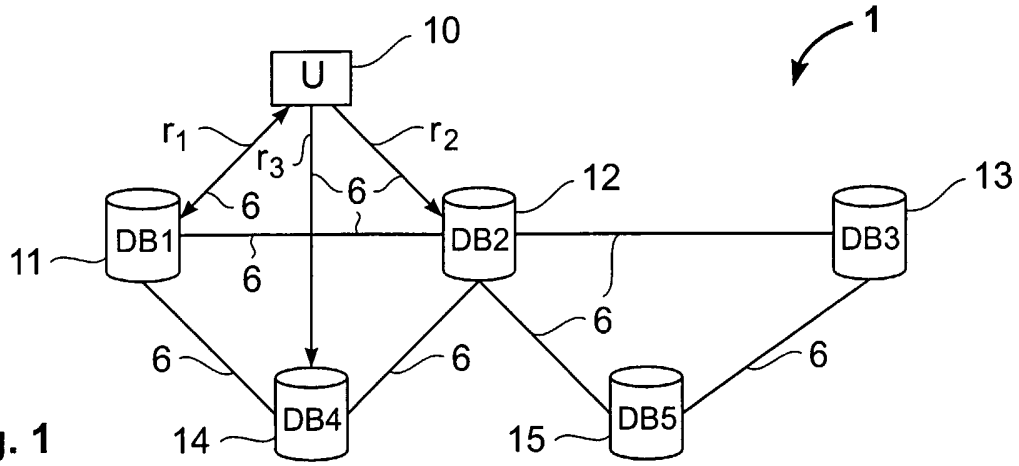

1
10
U
r1
r2
r3
6
12
13
DB1
DB2
DB3
11
6
6
6
6
6
14
DB4
DB5
15

Fig. 1

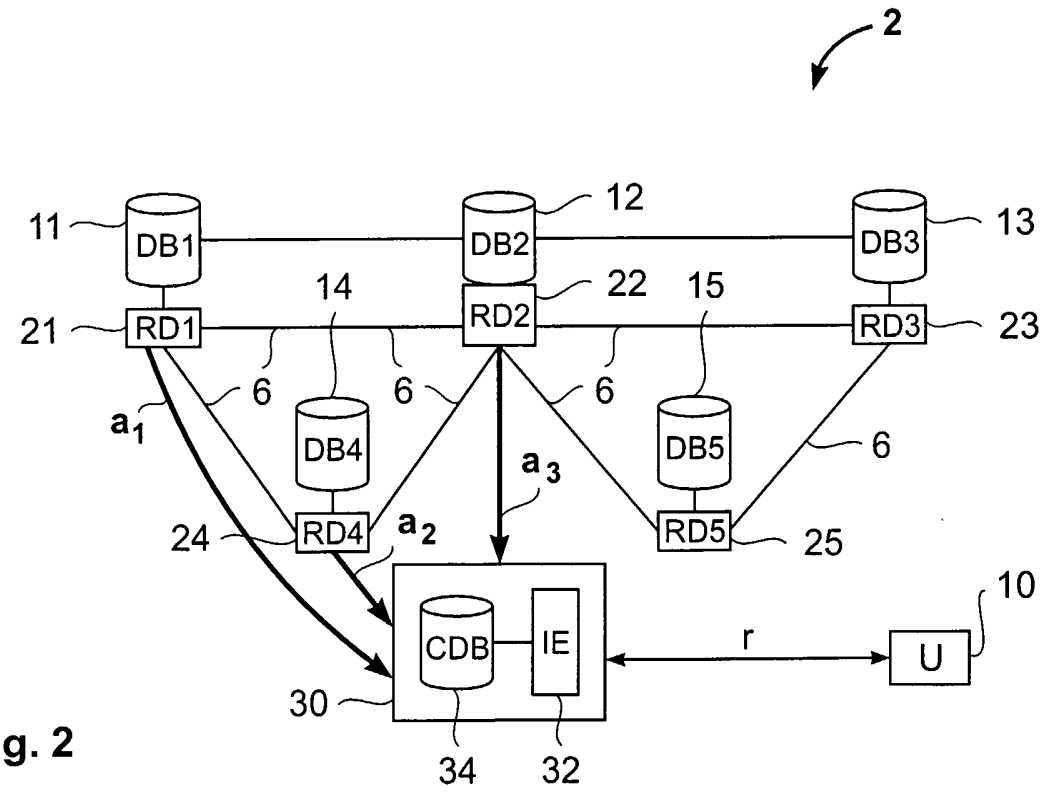

2
11
DB1
12
DB2
13
DB3
14
22
15
21
RD1
RD2
RD3
23
6
6
6
6
a1
DB4
DB5
a3
24
RD4
a2
RD5
25
CDB
IE
r
U
10
30
34
32

Fig. 2

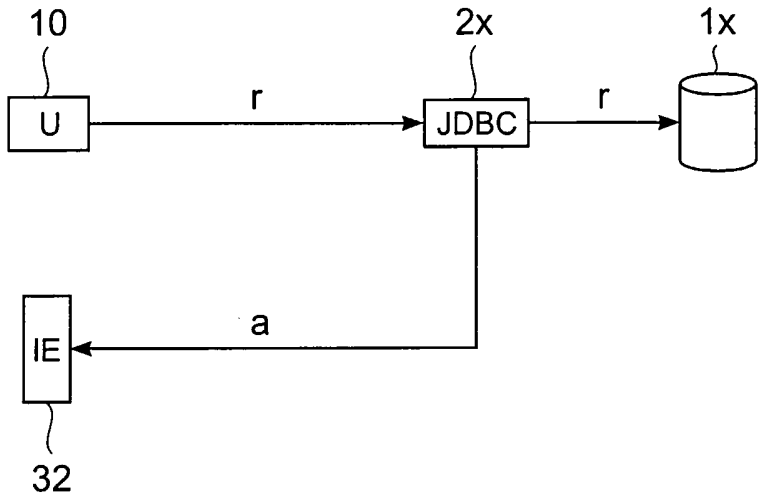


Fig. 3A

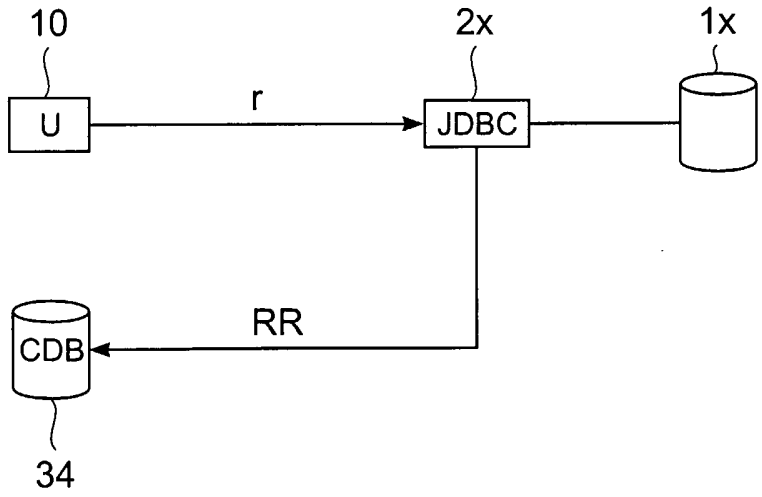


Fig. 3B

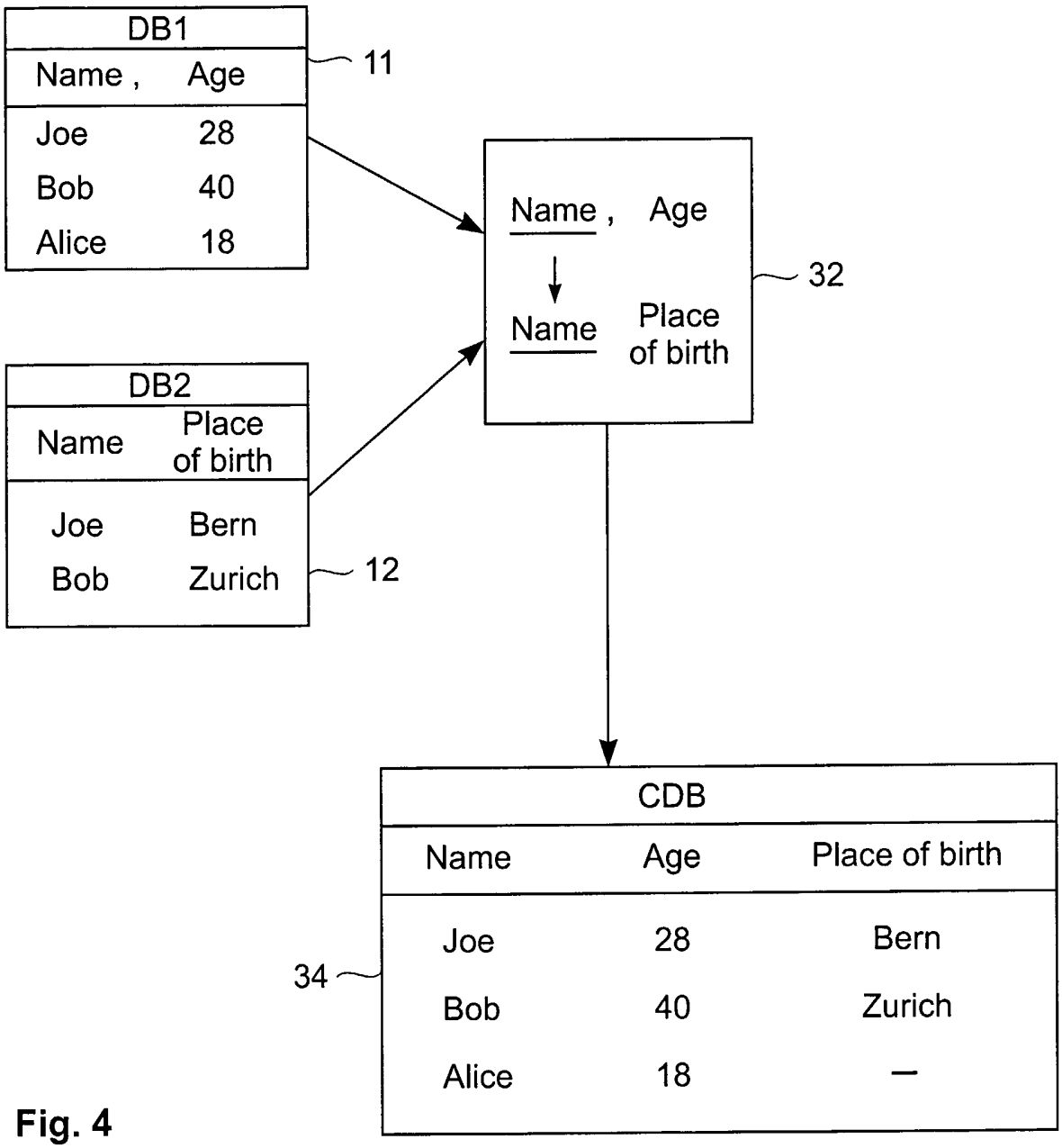

DB1
Name , Age
Joe 28
Bob 40
Alice 18
11
DB2
Name Place of birth
Joe Bern
Bob Zurich
12
Name , Age
Name Place of birth
32
CDB
Name Age Place of birth
Joe 28 Bern
Bob 40 Zurich
Alice 18 —
34

Fig. 4

INTERNATIONAL SEARCH REPORT

National Application No
. . T/IB2004/000669

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2002/107871 A1 (OLIVER WILLIAM J ET AL) 8 August 2002 (2002-08-08) paragraph '0004! - paragraph '0006! paragraph '0021! paragraph '0026! - paragraph '0028! paragraph '0033! - paragraph '0034! paragraph '0038!	1-13
A	US 6 151 608 A (ABRAMS HELENE G) 21 November 2000 (2000-11-21) the whole document	1-13
A	US 5 710 917 A (GODOY GLENN CARROLL ET AL) 20 January 1998 (1998-01-20) cited in the application the whole document	1-13

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 September 2004	13/09/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	DE CASTRO PALOMARES

INTERNATIONAL SEARCH REPORT

Information on patent family members

national Application No
. .T/IB2004/000669

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2002107871 A1	08-08-2002	NONE	
US 6151608 A	21-11-2000	AU 3475199 A	25-10-1999
		WO 9952047 A1	14-10-1999
US 5710917 A	20-01-1998	NONE	