



(19) **United States**

(12) **Patent Application Publication**
Parkkinen et al.

(10) **Pub. No.: US 2014/0136510 A1**

(43) **Pub. Date: May 15, 2014**

(54) **HYBRID TABLE IMPLEMENTATION BY USING BUFFER POOL AS PERMANENT IN-MEMORY STORAGE FOR MEMORY-RESIDENT DATA**

(22) Filed: **Nov. 13, 2012**

Publication Classification

(71) Applicant: **INTERNATIONAL BUSINESS MACHINES CORPORATION,**
Armonk, NY (US)

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(52) **U.S. Cl.**
USPC **707/711; 707/E17.108**

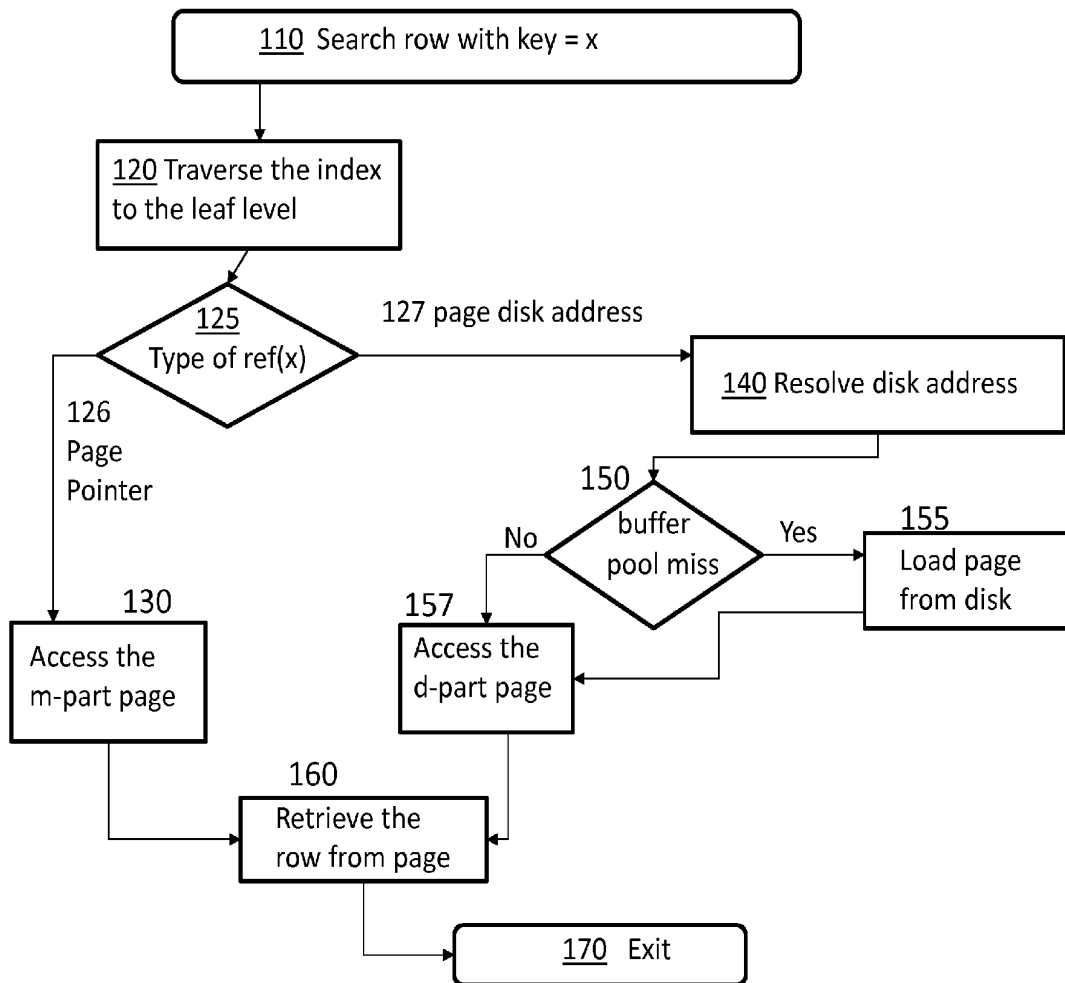
(72) Inventors: **Jarmo Parkkinen,** Helsinki (FI); **Vilho T. Raatikka,** Helsinki (FI); **Jarmo K. Ruuth,** Espoo (FI); **Petri U. Soini,** Vantaa (FI); **Antoni Wolski,** Kirkkonummi (FI)

(57) **ABSTRACT**

Embodiments relate to a method, system, and computer program product for data management. An aspect includes a search index that is created for a database. The search index has a reference value of a first or a second type associating data to a storage location in a first memory or a second memory respectively. Responsive to a data retrieval request, data is accessed using the reference value from the first or the second memory. If there is a data migration between the first and second memory storage, the reference values are then recalculated and altered in the search index accordingly.

(73) Assignee: **INTERNATIONAL BUSINESS MACHINES CORPORATION,**
Armonk, NY (US)

(21) Appl. No.: **13/675,634**



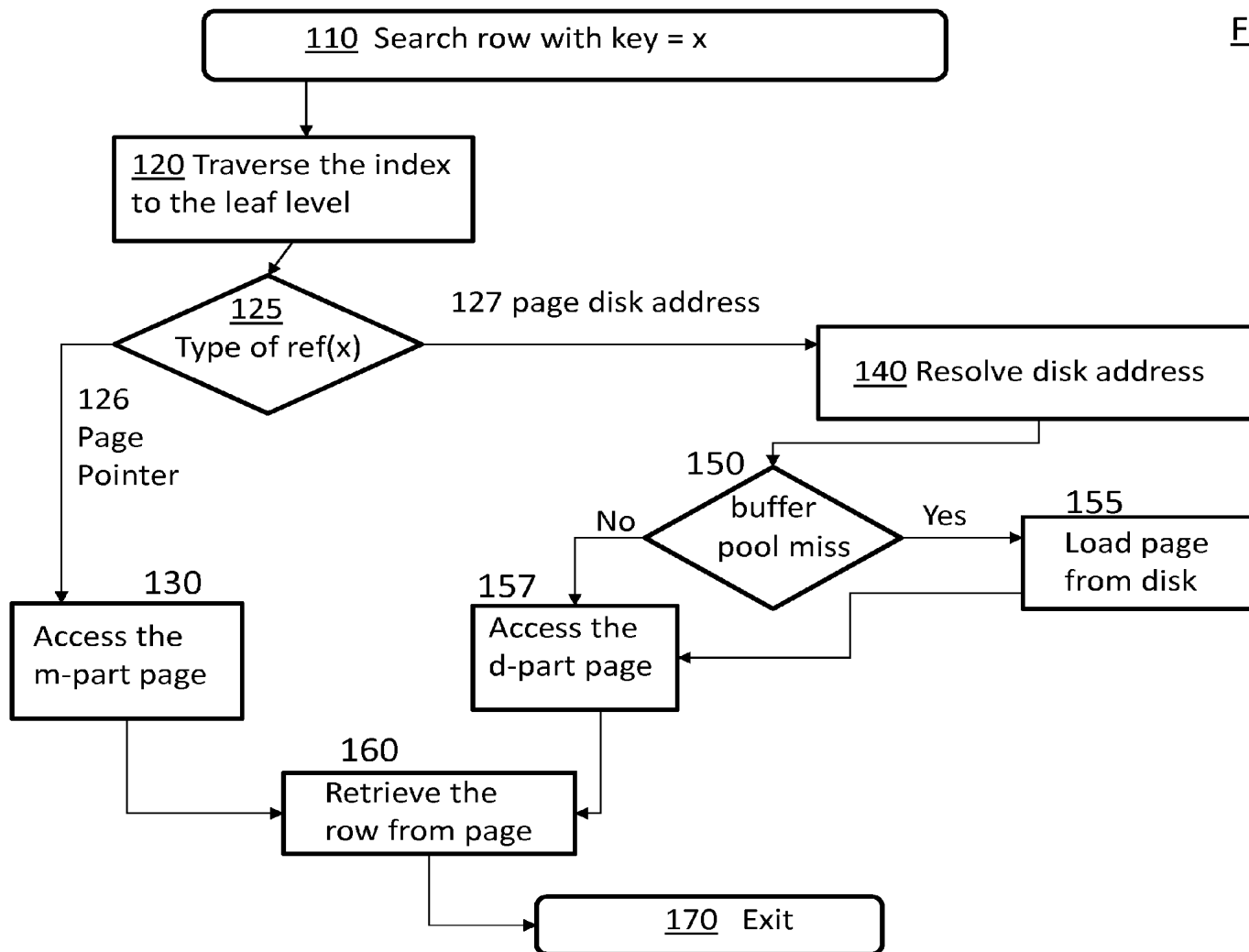


FIG. 1

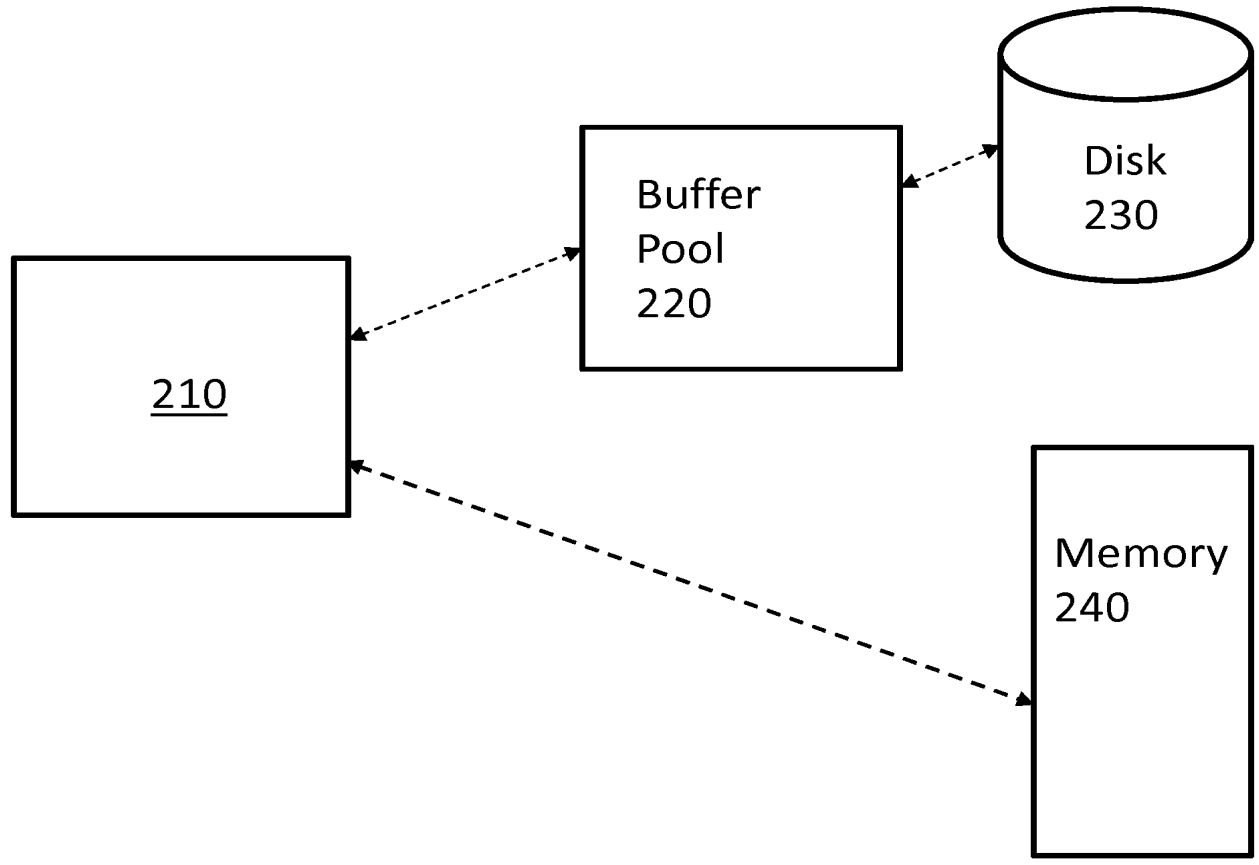


FIG. 2

**HYBRID TABLE IMPLEMENTATION BY
USING BUFFER POOL AS PERMANENT
IN-MEMORY STORAGE FOR
MEMORY-RESIDENT DATA**

BACKGROUND

[0001] The present disclosure relates generally to field of database search and storage; and more specifically to a hybrid table implementation by using buffer pool as permanent in-memory storage for memory resident-data.

[0002] Many large databases are primarily stored on a disk. These disk based databases, often use a buffer to retrieve newly accessed data to increase performance. Disk-based databases often share buffer space to optimize space and increase performance of the system. However, the pooling of the buffer creates performance bottlenecks when sending or receiving data, which are often caused by the disk input-output (I/O) requirements.

[0003] To reduce such bottlenecks and avoid the performance issues, in-memory databases are sometimes used. In an in-memory database, the primary location of data is in the physical or permanent memory. Most in-memory databases are characterized by memory optimized data structures and access methods. Performance is greatly enhanced by sorting, storing and retrieving specified data entirely from the in-memory, rather than going to disk based. Using an in-memory database allows access requests to be executed without dealing with the issues of disk I/O bottlenecks provided in the code path. However, using an in-memory database scheme has many trade-offs as the storage cost per data unit is higher for the in-memory database than for a disk memory scheme. This is because a less costly hard disk can be substituted for more memory. In addition, random access memories (RAM) chips used for in-memory database schemes cannot approach the density of a small hard drive. Another issue with in-memory databases is that in some applications, mining on large databases cannot be held into in-memory storage due to space constraints. One solution is to use a hybrid search index of databases which uses both "d" and "m" types of databases.

BRIEF SUMMARY

[0004] Embodiments include a method, system, and computer program product for data management. In one embodiment, a search index is created for a database and the search index has a reference value of a first or a second type associating data to a storage location in a first memory or a second memory respectively. Responsive to a data retrieval request, data is accessed using the reference value from the first or the second memory. If data is migrated between the first and second memories, the search index reference values are recalculated and altered.

[0005] Additional features and advantages are realized through the techniques of the present disclosure. Other embodiments and aspects of the disclosure are described in detail herein. For a better understanding of the disclosure with the advantages and the features, refer to the description and to the drawings.

**BRIEF DESCRIPTION OF THE SEVERAL
VIEWS OF THE DRAWINGS**

[0006] The subject matter which is regarded as the invention is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing

and other features, and advantages of the disclosure are apparent from the following detailed description taken in conjunction with the accompanying drawings in which:

[0007] FIG. 1 is a flowchart illustration in accordance with an embodiment; and

[0008] FIG. 2 depicts a block diagram illustrating memory access to data and a buffer, in accordance with an embodiment.

DETAILED DESCRIPTION

[0009] Hybrid search indexing of databases provides flexibility. A hybrid search index database both uses in-memory and disk-based databases. The distinction between the two types of data location is made at the table level. Tables can be designated as an in-memory table (m-table) where the entire contents of that table are stored in memory or alternatively tables can be disk-based (d-table) where the data is stored primarily on disk. Using both in-memory and disk storage achieves a balance between performance and cost.

[0010] In most databases, tables are set up and used for data retrieval. A table is a set of data elements or values that is organized into horizontal rows and vertical columns, having a plurality of cells where a row and a column intersect. Conventionally, a table has a specified number of columns, but any number of rows. Each row is identified by the values appearing in a particular column subset which has been identified as a unique key index. To improve the speed of data retrieval operations, a database index is used. Indices can be created using one or more columns of a database table, providing the basis for both rapid random lookups and efficient access of ordered records.

[0011] The dichotomy of using two different database types in a single setting poses when using hybrid search indexing still poses a challenge because these databases have different requirements. One conceivable solution is to provide a scheme where a new index is layered on top of one another (separate m-part and d-part indexes). But this solution can also be inefficient as it forces a continuous switching between the m-part and the d-parts of two separate tables when data is being accessed. This can become both resource intensive and time consuming. In addition, such an index may have footprint requirements that can prove to be too large for available storage in most memories.

[0012] FIG. 1 shows a flowchart implementation of one embodiment where a hybrid database is used that optimizes resource and search time. As shown in FIG. 1, while two types of memories are used, such as in one embodiment a disk storage memory and an in-memory, constant switching is not necessary as the combination of using a reference value and a buffer pool allow identification of storage of data in the disk storage or in-memory storage. The access to the buffer pool and to the disk and/or in-memory storage can be understood better by reference to FIG. 2.

[0013] FIG. 2 illustrates a block diagram as per one embodiment that can be used in accordance with the flowchart of FIG. 1. In FIG. 2, memory access to the storage database is optimized via the use of a buffer 220 which is in processing communication with an in-memory or permanent memory storage 240 and a disk memory storage 230. In the embodiment shown in FIG. 2, the in-memory or permanent memory is in processing communication with the buffer 220 via a search engine 210, but in alternate embodiments direct processing access can also be provided. It should be noted that permanent, physical and in-memory storages are used

interchangeably to mean the same type of memory storage. In one embodiment, memory access can be achieved to both memory storages via the search engine **210** that can be implemented using one or more processors **215**. The digital devices using the search engine **210** can represent a variety of digital devices, such as but not limited to mobile devices, personal computers, servers or and may include or be in processing communication further with displays, printers or other components as can be appreciated by those skilled in the art. In other embodiments buffer **220** and disk storage **230** and in-memory storage **240** may be part of a unitary device and access may be performed directly to them with or without a separate search engine **210** such as directly by a processor or computer as can be appreciated by those skilled in the art.

[0014] Traditionally, a buffer is a region of a physical memory storage used to temporarily hold data while it is being moved from one place to another. In this way, a buffer can hold information about accessed data from disk storage but only a limited quantity of the disk storage data can be buffered this way. In FIG. 2, the buffer **220** is buffer pool and can include a buffer pool manager **221**, in one embodiment. An optimized access to the database is then achieved by using the buffer pool as a permanent storage for memory resident data while disk resident data is buffered in the regular way as appreciated by those skilled in the art.

[0015] Optimized access as discussed in conjunction with FIGS. 1 and 2, is enabled to be used by many types of search index structures as known to those skilled in the art. An example of this would be a tree structure, simulating a hierarchical tree with a set of linked nodes. A node may contain a value or condition, or represent a separate data structure or even be a tree of its own. Each node in a tree has zero or more child nodes, which are below it in the tree. A node that has a child is called the child's parent node (or ancestor node, or superior). A node has at most one parent. An internal node (also known as an inner node or branch node) is any node of a tree that has child nodes. Similarly, an external node, also known as an outer node, leaf node, or terminal node, is any node that does not have child nodes. Similarly, a tree structure with a binary or B-tree structure can also be used with the optimized access described in conjunction with FIGS. 1 and 2. A binary tree (B-tree) data structure keeps data sorted and allows searches, sequential access, insertions, and deletions in logarithmic time. A B-tree is often has more than two children and is optimized for systems that read and write large blocks of data.

[0016] For ease of understanding, the flowchart embodiment of FIG. 2, shows a tree structure with leaf nodes but as discussed other search structures can be used in alternate embodiments. Referring back to FIG. 1, when access to data is needed a search is initiated by first finding the database associated with that data. In the database search index, in one embodiment, this can start by identifying a location in the index such as a row and associated key.

[0017] Referring back to FIG. 1, in one embodiment a search index table is then set up so that if the data row (block **110**) is found to be in the m-part of the table, the page is accessed via a pointer (stored in the leaf node of said index tree and said data found in the d-part of the table is accessed by invoking a buffer pool manager to resolve disk address and initiate a page buffer load. This concept will now be discussed by examining the blocks of FIG. 1's flowchart in more detail.

[0018] Starting with block **110**, the search index table in this case is finding the entry in this case by searching a

particular row with a particular key which by way of example is chosen to be key=x (e.g., a string of numbers or characters). The search is performed to the leaf level of the index structure as shown in block **120**. As shown in block **125**, the reference to the key is then tested to determine whether the reference is a page pointer or a page disk address. If the reference is a page pointer, then the page pointer is used to access the m-part page which is in the in-memory storage **130** (**240** in FIG. 2). At block **160**, the row that contains the data matching the key used in the search is then retrieved from the accessed m-part page. On the other hand, if the reference (x) **125** is a page disk address **127**, then further processing can be performed to resolve the disk address **140**. In addition, at block **150** the page disk address is tested to determine whether the page referenced thereby is present in the buffer pool. Thus at **150**, if there is no "buffer pool miss" (i.e. the data is present in the buffer pool), then as shown at **157** the d-part page corresponding to the page disk address is accessed from the buffer pool. Otherwise, if there is a buffer pool miss at **150**, then as shown at **155** the referenced page is accessed from the disk memory storage **230** (FIG. 2). A buffer pool manager is then invoked, in one embodiment, to transfer a chunk of data containing referenced page from disk memory storage into the buffer pool. In either case, as shown at **160**, the row that contains the data matching the key used in the search is then achieved from the accessed d-part page.

[0019] The optimized access as discussed in conjunction with FIG. 2, does not necessitate the need for a large memory footprint. Traditionally, m-table indexes are dense which means the databases is structured such that there is a file with pairs of keys and pointers for every record in the data file. In other words, every key in this file is associated with a particular pointer to a record in the sorted data file. In clustered indices with duplicate keys, the dense index points to the first record with that key. Most often there is a unique key and a reference for each row in the index. On the other hand, a d-table index is often sparse and structured so that a file is provided with pairs of keys and pointers for every block in the data file. Every key in this file is associated with a particular pointer to the block in the sorted data file. In a sparse index, a range of key values is mapped into a single data page. That is possible because the key values are clustered (adjacent values are located in the same page).

[0020] Consequently, it is much more advantageous in terms of footprint size to have a sparse database because a dense indexing of all the rows of a hybrid table would require too much space which is the challenge for keeping a single hybrid table. Moreover, with a single hybrid table, a typical page buffer pool solution has to be still retained for efficient loading of d-part data blocks into memory, for processing. Another challenge has been to facilitate easy migration of data between the "m" and "d" parts. However, a single table index that is efficient in accessing the m-parts as memory index and conserves space by including a sparse index, is not sufficient because data migrations between the m-part and the d-part can still be resource-consuming. This is because such migration requires copying the data to be physically transferred from the m-part to the d-part. The need for data migration normally results in data aging which means dealing with older data that is not used frequently and should be migrated to a slower medium.

[0021] In the optimized access as discussed in conjunction with FIG. 2, a sparse index can be used and the data can be migrated between the m-part and d-part by simply changing

the type of reference and recalculating the reference value. The contents of a page need not be changed. The migration can be done explicitly, in one embodiment, or be associated alternatively with a page replacement mechanism like an LRU (least recently used) queue. The m-part pages not used recently can then be changed to d-part change pages, which effectively inflicts the migration. This provides for an easy data migration between the m-part and the d-part. In addition, the efficiency of the m-table row access is preserved and the table row is accessed via an in-memory index and memory pointers. The access to an m-part row can be more efficient than to a d-part row because in this case the buffer pool manager has to be invoked to resolve the disk page address. The large capacity of the d-part is preserved by allowing the buffer pool manager to translate the disk page addresses to page hits and misses and act appropriately upon them. This addresses many of problems associated with use of index tables that are too large to fit into memory and inflexible and costly to search as discussed, especially in instances where data cannot be fit into memory itself and must be split into separate tables of different types.

[0022] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the disclosure. As used herein, the singular forms “a”, “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprises” and/or “comprising,” when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0023] The corresponding structures, materials, acts, and equivalents of all means or step plus function elements in the claims below are intended to include any structure, material, or act for performing the function in combination with other claimed elements as specifically claimed. The description of the present disclosure has been presented for purposes of illustration and description, but is not intended to be exhaustive or limited to the disclosure in the form disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the disclosure. The embodiments were chosen and described in order to best explain the principles of the disclosure and the practical application, and to enable others of ordinary skill in the art to understand the disclosure for various embodiments with various modifications as are suited to the particular use contemplated.

[0024] Further, as will be appreciated by one skilled in the art, aspects of the present disclosure may be embodied as a system, method, or computer program product. Accordingly, aspects of the present disclosure may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a “circuit,” “module” or “system.” Furthermore, aspects of the present disclosure may take the form of a computer program product embodied in one or more computer readable medium(s) having computer readable program code embodied thereon.

[0025] Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer

readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium may be any tangible medium that can contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device.

[0026] A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device.

[0027] Program code embodied on a computer readable medium may be transmitted using any appropriate medium, including but not limited to wireless, wireline, optical fiber cable, RF, etc., or any suitable combination of the foregoing.

[0028] Computer program code for carrying out operations for aspects of the present disclosure may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the “C” programming language or similar programming languages. The program code may execute entirely on the user’s computer, partly on the user’s computer, as a stand-alone software package, partly on the user’s computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user’s computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

[0029] Aspects of the present disclosure are described above with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products according to embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data pro-

cessing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0030] These computer program instructions may also be stored in a computer readable medium that can direct a computer, other programmable data processing apparatus, or other devices to function in a particular manner, such that the instructions stored in the computer readable medium produce an article of manufacture including instructions which implement the function/act specified in the flowchart and/or block diagram block or blocks.

[0031] The computer program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other devices to cause a series of operational steps to be performed on the computer, other programmable apparatus or other devices to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide processes for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

[0032] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of code, which comprises one or more executable instructions for implementing the specified logical function (s). It should also be noted that, in some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts, or combinations of special purpose hardware and computer instructions.

What is claimed is:

1. A data management method, comprising:
 - creating a search index for a database, said search index having a reference value of a first or a second type associating data to a storage location in a first memory or a second memory respectively;
 - responsive to a data retrieval request, accessing data using said reference value from said first or said second memory;
 - recalculating and altering said search index reference values based on data being migrated between said first and second memories.
2. The method of claim 1, wherein said first memory is in-memory storage and said second memory is disk memory storage.
3. The method of claim 2, further comprising using said reference value of said second type to access said particular data from a buffer pool if it is determined that said particular data is accessible from said buffer pool.
4. The method of claim 3, wherein said search index is a tree structured index.
5. The method of claim 4, wherein said tree has leaf nodes.

6. The method of claim 5, wherein said reference value of said first type is a page pointer to in-memory storage and said reference value of said second type is a page disk address.

7. The method of claim 6, wherein data access is provided using said page pointer or said page disk address according to said reference value provided.

8. The method of claim 5, wherein said leaf node indexing reference value for m-data in said permanent memory is a pointer to a page in said buffer and said leaf node indexing reference value for d-data is a disk address.

9. The method of claim 5, wherein said search index is a binary (B-tree) index structure.

10. The method of claim 6, wherein said index has a plurality of rows and columns and said rows have associated keys.

11. The method of claim 10, wherein said ordering of keys correspond to said ordering of rows.

12. The method of claim 8, wherein said migration from one component to another can be accomplished on a page replacement mechanism and said page is being generated by said buffer.

13. The method of claim 12, wherein said migration from one component to another can be accomplished on a least recently used basis.

14. The method of claim 8, wherein said migration from one component to another can be accomplished on an explicit basis.

15. A computer program product for data management, the computer program product comprising a computer readable storage medium having computer readable program code embodied therewith, said computer readable program code being executable by a computer to:

- create a search index for a database, said search index having a reference value of a first or a second type associating data to a storage location in a first memory or a second memory respectively;
- responsive to a data retrieval request, access data using said reference value from said first or said second memory;
- recalculate and alter said search index reference values based on data being migrated between said first and second memories.

16. The computer program product of claim 15, wherein said index has a plurality or rows and columns and said index has associated keys corresponding to each row.

17. The computing program product of claim 16, wherein indexing granularity is a page and said index is enabled to have an additional on-page access.

18. The computer product of claim 17, wherein said first memory is in-memory storage and said second memory is disk memory storage.

19. The computer program product of claim 17, wherein said index is a sparse index and said d and m pages are identical in structure except for content.

20. A system comprising:
 - an in-memory storage;
 - a disk memory storage in processing communication said in-memory storage;
 - a buffer in processing communication with said in-memory and disk memory storage, said buffer including a buffer pool manager for handling buffer pool;
 - a database search engine having at least a processor for managing a data search index structure, said search

index having reference values of a first or a second type such associating data to said in-memory and disk memory storage;
said reference value of first type being a page pointer to in-memory storage in said buffer pool and said reference value of second type being a page disk address to be processed by said buffer pool.

* * * * *