A Survey of Physical Design Techniques of Information Systems

Karim Ali and Sarah Nadi
{karim, snadi}@cs.uwaterloo.ca
David R. Cheriton School of Computer Science
University of Waterloo

Abstract

The abstract goes here.

1 Introduction

Relational Database Management Systems (RDBMS) have been the most popular type of Information Systems for decades. When we think of designing a database system, we usually think of which data we need to keep track of and how will we properly organize it into meaningful tables. However, this is not the only phase in database design. All of the designed tables need to be physically stored on disk. How these tables will be stored and accessed greatly affects the database performance. Physical database design is concerned with the actual storage of data on disk. This means how the files are organized, and how they are accessed. Data is stored in the form of records in files which reside on a physical disk. In this sense, we try to find the most efficient way possible to store and access the data such that queries are answered and executed efficiently. In the traditional relational database system, this involved selecting proper indexing techniques, partitioning the data for more efficient access, materializing the results of some queries, clustering and compressing the data, and many other techniques [1, 2]. The physical design aspects of relational databases have been studied in much detail. However, over time, different types of Information Systems have emerged that have different characteristics. For example, most relational databases have been disk-based where the primary copy of the data resides on disk. This is as opposed to Main Memory Relational Databases (MMDB) where the primary copy of the data resides in main memory and no disk operations are needed to query the data. Such differences pose new requirements for the physical design of these systems.

This survey presents the different physical design patterns that have been used in different types of information systems. The systems discussed in this survey are disk based relational databases (traditional database systems), main memory relational databases, data warehouses, and XML databases. The aim of this survey is not to explain the details of every data structure used in physical design, but rather to compare how the different data structures have been used in these systems. This includes explaining how alterations

or additions are made to data structures to fit the specific needs of every system. Using the different physical design methods used in disk based RDBMSs, we explore how these methods apply to the other systems and where the differences and similarities lie.

The rest of the survey is organized as follows. Section 2 first explains different elements of physical database design used in traditional disk-based databases. Section 3 then discusses how each of the three other systems uses these elements, as applicable. Sections 3.1, 3.2, 3.3, discuss Main Memory Databases, XML Databases, and Data Warehouses respectively. Throughout these sections, comparisons are made as to how the different elements have been modified to fit the needs of each system. Section 4 mentions some of the open problems of physical database design, and suggests possible solutions. Finally Section 5 summarizes and concludes this survey.

2 Elements of Physical Database Design

Relational Databases were first introduced by Codd [3]. Disk based Relational Database Management Systems (RDBMS) are the most common and popular type of Information Systems. Since they have been extensively researched and have been around for many decades, we use them as the basis of our comparison. In this section, we describe the different elements of physical design in RDBMS. These elements will also be examined in the other systems to see how they apply to them.

2.1 Index Structures

The first physical database design decision to be considered is the choice of indexes to be implemented. The concept of indices has been around for a long time. Just like and index at the end of the book is provided to help find certain content faster, indices in an information system help find content in tables faster. The main idea of an index is to have each record in a table have a unique identifier (primary key) and then organize these identifiers in a certain way that allows for fast access of a specific record. Of course, indexing can be done on any column in the database and not necessarily the primary key. There are many type of indexes that have been used in information systems.

Btrees (short for Balanced Trees) are the most used database index structure. The Btree is essentially similar to the traditional Binary Search Tree, but instead of having one value per node, a B-tee can have many values per node [4]. B-trees are suitable for relational databases since the cost of retrieval in a B-tree is at most proportional to: $log_d \frac{n+1}{2}$. The cost of insertion and deletion is at most proportional to $log_d n$ due to the possibility of progressing back up the tree to balance it after an insertion or a deletion. Although B-trees do well in retrieval, deletion and insertion, they do not perform well in sequential search.

Different variations of B-trees have been used. For example, B+trees are now the main method of indexing in current disk-based Relational Databases [2] such as DB2, Oracle, and SQL Server. The main difference between the B+tree and the Btree is that only leaf nodes contain data pointers in a B+tree, and leaf nodes contain pointers to each other as shown in Figure **??**. In Prefix B+trees, instead of using the actual key value of a record, a prefix from the key value is used to decrease the storage size needed which will

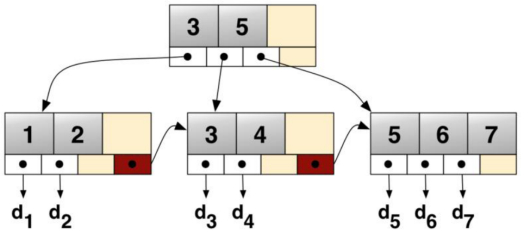

Figure 1: B+tree

in turn allow more records to be stored per node thus decreasing the height of the tree. B+trees have been successful in relational databases since data is ultimately recorded in files. Since B+trees have a high fanout, they allow less I/O operations to access a specific data record which is stored in a specific file.

Of course, there are other indexing techniques used such as hash tables, and bitmap indices. EXPLAAAAAAAAIN

briefly discuss clustered vs. unclustered ?

2.2 Materialized Views

In a large database system, there are usually a few complex queries that require the joining of many large tables. If these queries are frequently run, then the database performance is likely to suffer, and users have to wait for a long time to get the results back. If we know these queries beforehand, it makes sense to simply store the results of the queries on disk instead of recomputing them each time. Oracle 8i (SOURCE) introduced materialized views to fit this purpose. By precalculating the results of a complex query, and storing them in a table on disk, the new table with the results is very likely to be much smaller than the original tables decreasing I/O access costs and thus increasing performance [2, 5]. Once a materialized view exists, the user can either explicitly query the materialized view, or enter the original query, and have the query optimizer rewrite the query to use the materialized views instead of the original tables. Many query optimization techniques have relied on rewriting queries using views (E.g. [6, 7, 8, 9]), but discussing the details of the rewriting is beyond the scope of this survey.

Despite the fact that materialized views provide a significant improvement in performance, we cannot simply create a materialized view for each common query. To begin with, materialized views consume disk storage which might be limited. The second, and main, problem with materialized views is their maintenance. Ensuring that the views have the most up to date data is tricky, and may outset the benefit of using materialized views if it is not properly designed. Finally, having several materialized views may increase the cost of searching for the appropriate view to use during the query optimization stage. Data warehouses heavily rely on materialized views since they have many queries with several joins. This will be discussed in details in Section 3.3. There are many variations of B-trees. In relational databases, the most popular index is the $B^+ - tree$. The $B^+ - tree$ is a variant of the B tree which is becoming the main indexing method supported by many databases such as DB2, Oracle, and SQL Server citelightstone2007physical. The main difference between the $B^+ - trees$ and the B-tree is that only leaf nodes contain data pointers in a $B^+ - tree$. (COMPARE COST HERE). In Prefix $B^+ - trees$, instead

	Range Partitioning	List Partitioning	Hash Partitioning	Composite Partitioning
How?	If key falls in range	List of keys	Hash function	Hybrid
Why?	Disk limits	Group point values	Range/List are N/A	According to needs
Benefits	Partition elimination, Improved administration, Fast roll-in/out, and Implicit clustering			
Example	• Zip code • Time range	• Regions • States	• EmployeeId • ProductId	Hash then List
RDBMS	MySQL, DB2, Oracle, PostgreSQL, SQL Server	MySQL, Oracle, PostgreSQL, SQL Server	MySQL, Oracle	MySQL, Oracle

Table 1: Types of horizontal partitioning

of using the actual key value of a record, a prefix from the key value is used to decrease the storage size needed which will in turn allow more records to be stored per node thus decreasing the height of the tree.

The B+tree is the main indexing method used in current relational databases [2]. B+trees have been successful in relational databases since data is ultimately recorded in files. Since B+trees have a high fanout, they allow less I/O operations to access a specific data record which is stored in a specific file.

2.3 Partitioning

Partitioning is an important aspect of physical design as it reduces table scanning time. There are two main categories of partitioning: horizontal partitioning and vertical partitioning. Horizontal partitioning divides the table into sets of table rows where each row or record still has all the attributes of the table. For example, dividing the data by date where all data dating less than year 2000 lies in one partition while all data dating more lies in another. On the other hand, vertical partitioning reduces the width of the dataset by storing some attributes in one table, and some in another. Usually, less frequently queried columns are partitioned together to reduce the size of the table.

The main advantage of any type of partitioning is to reduce the amount of time it takes to scan a table which in turn improves performance. For example, if a table has 1000 records, and a query is only using the first 20 records then having these 20 records in a partition containing 50 records will save time as opposed to examining all the 1000 records. Horizontal partitioning is more widely used, and is implemented in all current DBMSes. Table 1 shows different categories of horizontal partitioning, and the situations they are suited to. Combining partitioning and indexing can often lead to better performance. For example, if List partitioning is used to divide a table by regions and then a bitmap index is used on that table, finding records in the same region or state can be much faster. Finally, partitioning the data into different physical devices to achieve parallelism when accessing it can greatly increase the performance of the database.

2.4 Clustering

When only records satisfying certain criteria are need from a table, it is often unnecessary to scan the whole table. For example, if only records that occurred in January 2010 are needed, then querying all the table would be an extremely unnecessary overhead. This is where clustering comes in. Clustering reflects how the records are physically located together on disk. Records that are likely to be queried together will be stored physically together to avoid querying the whole table. For the previously mentioned example, the data can be clustered by month, such that all the records of January 2010 are physically in the same page, for example. In this case, only this page will need to be read. As people realized the potential of clustering, multidimensional clustering was introduced where data is clustered based on different criteria at the same time (TALK MORE).

2.5 Other Physical Design Techniques

There are other physical design methods to improve the performance of the information systems [2]. For example, data compression can be used to make more data fit into a fixed amount of disk space such that it can be access faster. However, compression is only beneficial if the cost of executing the compression algorithm is lower than the I/O cost of accessing the data. Otherwise, compressing the data must just add an overhead without improving performance. Data striping is used to distribute data that is accessed together across multiple disks. This allows the data to be retrieved faster through parallelism. Striping has the same goals as partitioning. However, striping is more geared towards the hardware level which is beyond the control of the database while partitioning is within the control of the database. Database reliability is also improved through mirroring which involves duplicating the data on multiple disks. Finally, refining the global schema to reflect query and transaction requirements is also sometimes used. This is called denormalization which can be thought of as manual clustering of the data to improve performance.

3 Physical Design of Different Information Systems

Given the different physical design methods explained in the previous section, we now look at how these methods apply to different types of information systems. Mainly, how they apply to Main Memory Relational Database Systems (MMDB), Data Warehouses, and XML databases. We mainly focus on the first four techniques discussed in Section 2 as these are the main design decisions involved in physical design.

3.1 Main Memory Relational Database System (MMDB)

A Main Memory Database (MMDB) (sometimes referred to as in-memory database) is one where the data resides in the main memory of the system rather than on a disk [10]. It should be noted that this is different from caching. Disk based database systems use the main memory to cache query results. However, MMDB's primary copy of the data resides in main memory. MMDB have many implementation challenges that have been addressed throughout the research community. Accessing data in main memory is faster

Physical Design Element	**Desired Features**	**Methods/Structures Used**
Index Structures	• No need to store actual values in index • Larger node size that is aligned with cache line size for cache consciousness	• B Tree • B+ Tree • CSB+ Tree • pB+ Tree • T Tree • CST Tree • CSS Tree
Materialized Views	• Not needed since processing and memory access is cheap	N/A
Partitioning	• Only needed in secondary storage to speed up reloading in case of a crash	• Horizontal Partitioning • Single Vertical Partitioning
Clustering	• Not needed since sequential random or dispersed main memory access is not more expensive than sequential access	N/A

Table 2: Summary of physical design of Main Memory Databases (MMDB)

than disk, and so MMDBs usually have better performance that disk based systems. However, main memory is more expensive and the size is limited in comparison to disk. However, MMDBs are essential to use where some of the collected data cannot, or will not, be stored on disk in the first place. For example, phone switch data may be needed for reasoning with other stored data, but it is captured in real time and will not necessarily be stored on disk.

Garcia-Molina and Salem [10] mention some of the challenges involved. One is concurrency control. In disk based databases, locks are tracked through a hash table, but the actual objects on disk do not contain lock information. On the other hand, in MMDB, lock status is part of the object itself since it is cheap to keep a number of bits for that. The next challenge is access methods. In disk-based databases, B-Trees are used to index the data for faster access. B-Trees have a short bushy nature to try to decrease the height of the tree to decrease the number of I/O accesses. Since I/O access is not a problem in MMDB, longer tree structures are used since its cheap to access main memory. Another point is that data values do not need to be stored in the index itself since they will be stored in main memory anyways so there are no performance gains from storing them in the index itself. Table 2 summarizes the physical design elements of MMDBs.

3.1.1 Index Structures

Two considerations are taken when designing index structures for MMDB. The first is that the data resides in main memory and not on disk, and so many of the considerations taken for I/O operations is no longer there. Instead of focusing on disk access and disk storage, the data structures for main memory databases should focus on the efficient use of CPU cycles and memory [11]. The second is having index structures cache conscious. Cache memories improve performance by holding recently referenced data [12]. If the memory reference is found in the cache, then execution proceeds at processor speed. Otherwise, the data has to be fetched from main memory. With the big difference between processor speed and main memory speed, cache misses are still relatively expensive. Lots of research work has shown that cache performance is very critical in MMDBs [13, 14]. This is because the big difference between processor speed and main memory access speed. Thus, a cache miss is still relatively expensive in main memory databases. Accordingly, we will see that many of the index data structures proposed for main memory databases focused on being cache conscious or cache sensitive. In summary, a main memory index should be able to reduce overall computation time while using as little memory as possible [11]. It should also be noted that unlike disk based systems where it is advantageous to store actual attribute values in the index, main memory systems do not require that. Instead, pointers to the actual attribute values can be used. This, in turn, will reduce the size of the index which is preferable in this case.

In this section, we talk about three families of index structures used in MMDB: B Trees, T Trees, and Binary Search Trees. There are more recent index structures developed for MMDB such as J+ Trees [15], but which have not gained much popularity, and so we do not discuss them in our survey. Table 3 summarizes the features of the different index structures discussed in this section with respect to MMDB. The T Tree is the most used index for MMDB (E.g. MySQL CLuster). However, there are other indexes used as well. For example, IBM's solidDB uses Trie (or prefix tree) for its indexing. For their purposes, it served as a good index for main memory since it eliminates the need for many key comparisons [16].

B Trees

Although B Trees, explained in the previous section, are originally designed for disk based database systems, they were found to also be useful in MMDB [11]. In disk based systems, the B+ Tree is more used than the regular B Tree. However, in main memory databases, the B Tree would be more suitable from a storage perspective since there is no gain from keeping all the data in the leaves. In a main memory system, this would only waste space without improving performance. On the other hand, the B+ Tree uses multiple keys to search within a node. This means that if the node fits in a cache line, this cache load can satisfy more than more comparison leading to better cache utilization [14]. Therefore, B+ Trees in general have reasonable cache behavior.

However, the cache behavior of B+ Trees could still be improved. Therefore, Rao and Ross [17] propose Cache Sensitive B+-Trees ($CSB^+ - Trees$) [17] as an index structure for main memory. To do so, they eliminated most of the child pointers and had more keys in each node to improve locality and reduce tree height [15]. Since the number of cache misses in search operations is proportional to the height of the tree, $CSB^+ - trees$

Index Structure	Features suitable for MMDB	Cache Consciousness
B Trees	• Good storage utilization • Reasonably quick searching • Fast updating	• Reasonable cache behavior if node fits in cache line
B+ Trees	• Wastes space since all data is stored in the leaves • Reasonably quick searching • Fast updating	• Reasonable cache behavior if node fits in cache line
CSB+ Trees	• Same features of B+ Trees	• Good cache performance due to improved locality & lower tree height
pB+ Trees	• Same features of B+ Trees	• Improves cache miss performance by prefetching more data in each cache miss • Node size is also increased leading to the same benefits of CSB+ Trees
T Tree	• Contain pointers to data values instead of the values themselves leading to better space utilization • Node size is bigger leading to previous mentioned advantages	• Cache behavior was not considered at time of design
CST	• Binary search tree for maximum value of each node leads to faster search • More space efficient by removing pointers & using array storage. Index calculation used to allocate values	• Node size is aligned with cache line size to avoid misses
CSS Trees	• Fast traversal in $log_{m+1}n$ (m is the number of keys per node) • Mainly suitable for read environments	• Good cache behavior since m is chosen to fit in the cache line size

Table 3: Summary of index structures for Main Memory Databases (MMDB)

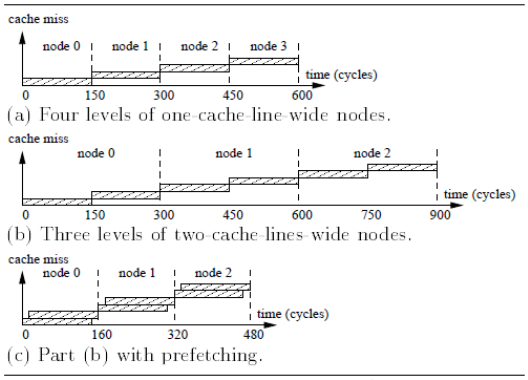

Figure 2: Performance of various B+ Tree searches where a cache miss to memory takes 150 cycles and a subsequent access can begin 10 cycles later [18]

have fewer cache misses, and thus better performance.

Chen et al. [18] propose using prefetching to improve the performance of B+ Trees. They call the new index structure prefetching B+ Tree (pB+ Tree). Since current computer systems can prefetch different data simultaneously, the authors take advantage of this. The idea here is to increase the node size of B+ Trees and employ prefectching at the same time. In a B+ Tree, every time we move down a level in the tree, a cache miss occurs. Accordingly, prefetching larger sizes nodes means that we can overlap multiple cache misses before they occur. This is because in one cache miss, extra data is retrieved in parallel with the originally requested data if it can be predicted sufficiently early. Figure 1 shows how prefetching retrieves the same data in less time. This method, however, does not eliminate the full cache miss latency with every level in the tree.

T Trees

Lehmen and Carey [11] introduce a new index structure, the T Tree, which combines the features from AVL Trees and B Trees that are suited to main memory. Figure 2 shows the proposed structure of the T Tree. They run some experiments on data residing in main memory to compare the performance of the T Tree to existing structures. All structures were modified to contain pointers to data values rather than the data values themselves. The structures compared included AVL Trees, simple arrays, B Trees, Chained Bucket Hashing, Extendible Hashing, Linear Hashing and Modified Linear Hashing. Their experiments showed that for unordered data, Modified Linear Hashing gave the best performance, and for ordered data, T Trees gave the best performance for a mix of searches, inserts and deletes. This is mainly because a T node contains many elements which results in good update and storage characteristics. The T Tree was the first index structure specifically designed for main memory databases.

Although T Trees have been specifically designed for MMDB, they did not consider the cache behavior at that time [14]. Although many keys are fit into one node, only the two end keys are used for comparison leading to low node utilization. Similarly, array binary search trees have the same problem.

T Trees were the main index structure used for main memory databases for some time after their proposal. However, it was discovered that B+-trees outperform T-Trees

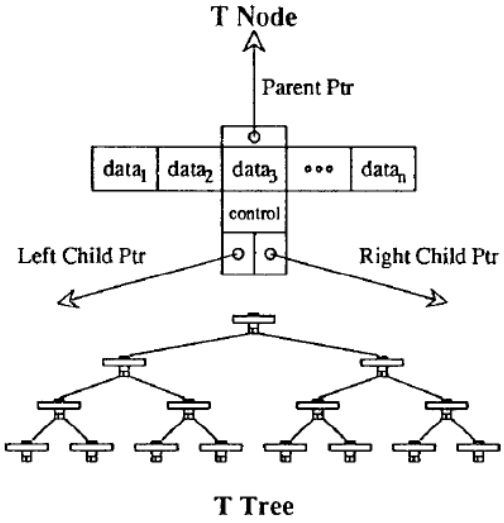

Figure 3: T Tree by Lehmen and Carey [11]

on modern processors because of the growth of cache miss latency relative to processor speed [14, 19]. This is because the height of the tree is high which makes the total number of memory accesses from the root to the leaf node higher. The other reason is that the node size is not aligned with the cache line size. Another problem with T-Trees is that there is a big waste of space in the cache because unnecessary data is brought into the cache. Additionally, record pointers stored in the tree take up a lot of space.

To resolve these issues, Lee et al. [19] propose the Cache Sensitive T-Trees (CST-Trees). This is achieved in the following ways. First, a binary search tree whose node values are the maximum key of each node in the T Tree is constructed. That way, for a given value, the node containing this value could be quickly located through searching the binary tree. Second, the need to store pointers is eliminated by storing the tree in an array and locating the necessary nodes through index calculation. Finally, node sizes are aligned with cache line size such that there are no cache misses when accessing data in each binary search tree in the array.

Cache Sensitive Search Trees (CSS-Trees)

Binary Search Trees have poor cache behavior when the array is much bigger than the cache. Rao and Ross [14] propose the Cache Sensitive Search Tree (CSS-Tree) to improve this cache behavior. To do that, the search tree is stored as a directory in an array. The number of keys per node is chosen such that the whole node fits in a cache line. This improves local searching within a node to just one cache miss. They also hard-code the traversal within a node such that finding the next node happens in $log_{m+1}n$ times rather than $log_2 n$ times in a binary search tree where m is the number of keys per node. However, CSS Trees are only suitable for read environments since they eliminate nearly all child pointers, thus removing the support for incremental update.

3.1.2 Materialized Views

From our literature survey, it seems that materialized views are not used in MMDB. This makes sense since the problems materialized views try to address are not present in MMDB. That is, the cost of computing complicated join queries is much less in MMDB due to the low cost of accessing the data in memory. Accordingly, the maintenance cost of materialized views in MMDB will outweigh their minimal benefit. However, this does not mean that they cannot be used at all in MMDBs. As with all other physical design decisions, it rather depend on the benefits of implementing the technique. Accordingly, if the data is very large and may benefit from using materialized views, there is no reason not to use them.

3.1.3 Partitioning

Partitioning is used in disk based systems to divide the data into the physical pages in a way that will yield the best performance. This could be through vertical partitioning or horizontal partitioning, as previously explained. In main memory databases, we do not have the issue of expensive disk access which makes partitioning in memory unnecessary as it will not save any costs. However, partitioning is needed in the secondary storage of the data on disk [20]. Disk storage is still used as a backup of main memory databases. When a crash occurs, a reload of the database from disk or archive memory takes place. In order to avoid page faults caused by referencing data that is still being reloaded, the reloading process must be designed to be efficient in loading the important data such that the database users are not affected. An efficient MMDB reload takes into consideration the number of I/Os for reload and the number of main memory references during transaction processing. Gruenwald and Eich [20, 21] show that horizontal partitioning is the most suitable when the database performs more modifications than tuple deletions and more selections that projections and joins. Single vertical partitioning is chosen in the other cases. They also show that if we just concentrate on the reload performance (i.e the number of I/Os for reload), then single vertical partitioning is always the best choice.

3.1.4 Clustering

Clustering in disk based systems is important because sequential I/O access is cheaper than random or dispersed I/O access. This is not the case with main memory, and so clustering is not needed in MMDB [10, 22]. Therefore, components of one object may be spread across memory without impacting performance.

3.2 XML Databases

XML (eXtensible Markup Language) is increasingly being used as a data exchange format. Accordingly, the ability to store and manage XML documents is needed. Salminen and Tompa [23] define an XML database as "a collection of XML documents and their parts, maintained by a system having capabilities to manage and control the collection itself and the information represented by that collection." Salminena and Tompa also highlight some of the requirements of an XML database system. These requirements include querying parts of an XML document, and transformations of XML documents.

The main challenge in XML databases is path navigation since it is a very expensive operation.

There are two ways in which an XML database system can be designed. The first is to have the database internally translate the XML document into relational tables (XML-enabled databases). The second is to have data structures that can persistently store XML documents (native XML databases). XML-enabled databases are suited for data-centric documents which have a regular structure without any mixed content [24]. On the other hand, for document-centric documents which have an irregular structure with lots of mixed content, having a native XML database is essential. In this section, we will discuss how the database's physical design needs to be adapted in both cases. Other challenges such as expressing and executing queries and updates are faced in XML database design, but these are beyond the scope of this survey.

Examples of XML-enabled databases include SQL Server 2005 and Oracle. Examples of native XML databases include Tamino [25], eXist [26], and TIMBER [27].

3.2.1 Index Structures

Despite the special nature of XML documents, the same index structures used in relational databases can still be utilized in XML databases. However, some adjustments may need to be made to fit the nature of the XML documents. For XML-enabled databases, XML documents are usually represented as relational tables, and then indexed similar to other tables. Pal et al. [28] describe how this is done in Microsoft SQL Server 2005. To start with, a new data type called 'XML' was introduced. This data type could contain values of complete XML documents, or just fragments of XML data. First, the different nodes in the XML data are labeled with the ORDPATH [29] mechanism. The XML data is then shredded into a relational table with five main columns: ORDPATH, TAG, NODE_TYPE, VALUE, and PATH_ID. This information is then stored in a $B^+ - tree$ as the primary index of the XML type. Additionally, a secondary index can be created on any of the columns of the primary index to speed things up.

For native XML databases such as eXist [26] or TIMBER [27], $B^+ - trees$ were still used to index the XML documents. However, in contrast to XML-enabled databases, the XML data is not first stripped into a relational table. eXist uses a number schema to assign unique identifiers to each node in the XML document. These unique number identifiers allow the determination of any node's parent, sibling or possible child nodes. Four index files are created for XML data where all the indexes are based on $B^+ - trees$. One index manages the collection hierarchy, one index collects nodes in a paged file and associates unique node identifiers to the actual nodes, one index indexes elements and attributes, and the last index keeps track of word occurrences and is used by the full text search extensions. Oracle uses a similar technique for indexing its XML content where an XMLIndex table is created for XML entries [30]. This table contains a unique identifier for the XPath leading to the node, the row id of the table used to store the XML data,an order key that identifies the hierarchal positioning of the node, a locater key used for fragment extraction, and the text value of the node.

TIMBER also uses a similar numbering schema. There are minor differences between the two numbering techniques, but the main idea is the same. The numbering is based on three values of a node: its start label, its end label, and its level (i.e. nested depth). On

the other hand, Sedna [31] uses a different reasoning to produce its numbering schema. Sedna assigns string identifiers to nodes such that they are lexically ordered according to the position of the node. Despite the differences in the ways the numbering is constructed, it seems that that no novel indexing structure is needed for XML databases. The problem is rather how to assign the nodes unique identifiers (the numbers or labels in the numbering schema) that can be used as identifiers in the index to reflect the XML document structure. Additionally, in native XML databases, documents are usually grouped into collections, and an index may be used to search within these collections. This is different from the node index which searches within a document.

Besides indexing the actual nodes, path indexing is also common in XML databases since usually queries search for a specific path which can be very expensive. This is done by clustering together all the IDs of nodes on a given path, and creating an index for them [32, 33]. Other techniques include that proposed by Cooper et al. [34] where they propose an indexing mechanism for paths called "Index Fabric". An Index Fabric is based on Patricia tries [35] which are used for string indexing, but is customized for disk-based systems. A path in an XML document is encoded into a unique string, and is then inserted into the fabric. The encoded string is usually very short, and so searching is relatively fast.

3.2.2 Materialized Views

The main idea behind using materialized views is to rewrite queries using these views so that they run more efficiently. This is very important in XML Databases since XQuery and XPath queries are complicated [36]. The semi-structured nature of XML data, and the expressiveness of most XML querying languages, view selection for optimization becomes a more complicated problem [37]. Therefore, the choice of which views to materialize becomes more complicated in XML database. Much work has been done on rewriting XML queries using views [36, 38, 39, 40]. However, this problem is not within the scope this survey.

3.2.3 Partitioning

In XML-enabled databases, inlining during the shredding of the data such that children nodes that occur only once are stored with the parent node produces better partitioned data [41]. Inlining shreds a document into sets of tables according to the document schema which leads to better performance because queries tend to access less data. Such inlining is similar to horizontal partitioning [42]. Additionally, it is common to horizontally partition the mapped relational tables such that nodes of the same type are stored together [43].

3.2.4 Clustering

Clustering in XML databases follows the heuristic that sub-elements are likely to be queries with an element [27]. Accordingly, elements and their subelements are clustered together. In general, storing the XML data in document order it thought to be the best technique for efficient querying. Additionally, Lian et al. [44] suggest clustering XML documents by structural similarity to produce more efficient access. This is done by representing the documents as a structure graph (s-graph), and then identifying clusters

based on similarity using a clustering algorithm. Nayak [45] also proposes a similar clustering mechanism for XML data. Her technique is also based on structural similarity, but it is more focused on the level structure of the XML documents.

3.3 Data Warehouses

A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data and decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions [46, 47]. More than half of IT executives named Data warehousing as the highest-priority post-millennium project for them [48]. The value of data warehousing for an organization depends on the organization's need for reliable, consolidated, unique and integrated reporting and analysis of its data, at different levels of aggregation.

Data warehousing has been shown useful in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis) [47].

Data warehouses tend to be extremely large, in fact it is quite possible for a data warehouse to store tens of petabytes of data while loading tens of terabytes of data everyday [49]. Datta et al. [50] notes that the information in a data warehouse is usually multidimensional in nature, requiring the capability to view the data from a variety of perspectives. Aggregated and summarized data become more crucial than detailed records in such environment. Therefore, the workloads are query intensive with mostly ad hoc, complex queries, often requiring computationally expensive operations such as scans, joins, and aggregation. Performing such operations on large amounts of data, like in the case of data warehousing, complicates the situation further. Moreover, the results have to be delivered interactively to the business analyst using the system.

Although data in a data warehouse is extracted and loaded from multiple on-line transaction processing (OLTP) data sources (including DB2, Oracle, IMS databases, and flat files) using Extract, Transfer, and Load (ETL) tools (see Figure 3), a data warehouse is usually maintained seperately from the organization's operational databases [48, 47]. This architecture can be justified owing to the fact that operational databases are finely tuned to satisfy known OLTP requirements and functionalities which are quite different from that of on-line analytical processing (OLAP) which is supported by data warehousing. Analyzing data for decision support usually requires consolidating data from many heterogeneous sources of varying quality, or use inconsistent representations, codes and formats. Moreover, understanding trends or making predictions requires historical data, whereas operational databases store only current data. In OLAP, there's a need to support multidimensional data models and operations which requires special data organization, access methods, and implementation methods, not generally provided by DBMSs targeted for OLTP [47]. Finally, in data warehousing, query throughput and response times are more important than transaction throughput which is the major performance metric for operational databases.

[51] ROLAP o Data is stored in relational databases. o support for multi-dimensional

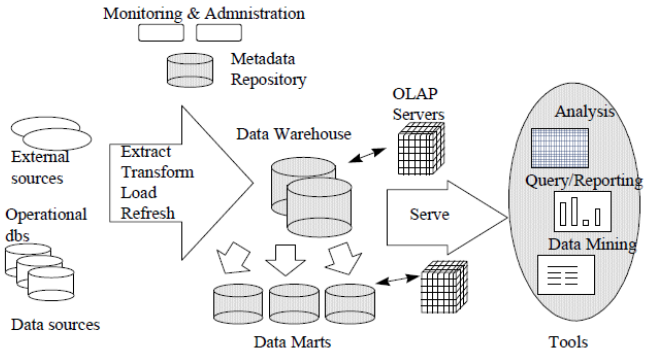

Figure 4: Data Warehousing Architecture [47]

view of data is achieved thru star scheme o Support extensions to SQL. o Efficiently implement the multidimensional data model and operations. o Ex: Microstrategy. MOLAP o Directly store multidimensional data in special data structures (arrays). o Implement the OLAP operations over these special data structures. o Snowflake Schema. [52] o Fact Constellations. o Ex: Essbase (Arbor), statistical DBMSes.

hierarchical data clusters [53].

star schema fact table

Typical Operations * Data Cleaning * Load * Refresh

Joins in data warehouses are usually expensive since the fact table (the largest one) participates in every join [50]

3.3.1 Index Structures

Tree-based Indexes B-tree (B+ trees but commonly referred to as B-trees) [54], R-tree [55, 51], K-D-B-tree [56], BV-tree [57], UB-tree [58]

Bitmapped Join Indexes [59]

Projection Indexes Involves positional indexing where tuples are accessed based on their ordinal position [50]

Bit-sliced Indexes

3.3.2 Partitioning

[49]

Vertical Partitioning

Horizontal Partitioning

3.3.3 Materialized Views

not very useful since most queries are ad-hoc [60] [61]

4 Open Problems in Physical Database Design

Indexing in the cloud [62].

Two-Tier Storage DBMS (using disk based and main memory dbs) [63].

5 Conclusions

References

[1] S. Finkelstein, M. Schkolnick, and P. Tiberio, "Physical database design for relational databases," *ACM Trans. Database Syst.*, vol. 13, no. 1, pp. 91–128, 1988.

[2] S. S. Lightstone, T. J. Teorey, and T. Nadeau, *Physical Database Design: the database professional's guide to exploiting indexes, views, storage, and more (The Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007.

[3] E. F. Codd, "A relational model of data for large shared data banks," pp. 5–15, 1988.

[4] D. Comer, "Ubiquitous b-tree," *ACM Comput. Surv.*, vol. 11, no. 2, pp. 121–137, 1979.

[5] S. Chaudhuri, "An overview of query optimization in relational systems," in *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. New York, NY, USA: ACM, 1998, pp. 34–43.

[6] A. Y. Levy, A. O. Mendelzon, and Y. Sagiv, "Answering queries using views (extended abstract)," in *PODS '95: Proceedings of the fourteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. New York, NY, USA: ACM, 1995, pp. 95–104.

[7] A. Gupta, V. Harinarayan, and D. Quass, "Aggregate-query processing in data warehousing environments," in *VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 358–369.

[8] J. Goldstein and P.-A. Larson, "Optimizing queries using materialized views: a practical, scalable solution," in *SIGMOD '01: Proceedings of the 2001 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2001, pp. 331–342.

[9] S. Abiteboul and O. M. Duschka, "Complexity of answering queries using materialized views," in *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. New York, NY, USA: ACM, 1998, pp. 254–263.

[10] H. Garcia-Molina and K. Salem, "Main memory database systems: An overview," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, pp. 509–516, 1992.

[11] T. J. Lehman and M. J. Carey, "A study of index structures for main memory database management systems," in *VLDB '86: Proceedings of the 12th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1986, pp. 294–303.

[12] A. J. Smith, "Cache memories," *ACM Comput. Surv.*, vol. 14, no. 3, pp. 473–530, 1982.

[13] P. A. Boncz, S. Manegold, and M. L. Kersten, "Database architecture optimized for the new bottleneck: Memory access," in *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 54–65.

[14] J. Rao and K. A. Ross, "Cache conscious indexing for decision-support in main memory," in *VLDB '99: Proceedings of the 25th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 78–89.

[15] H. Luan, X. Du, and S. Wang, "Prefetching $J^+ - Tree$: A Cache-Optimized Main Memory Database Index Structure," *Journal of Computer Science and Technology*, vol. 24, no. 4, pp. 687–707, 2009.

[16] S. H. Anotoni Walski, "solidDB and the secrets of speed. How the IBM in-memory database redefines high performance," Tech. Rep., 2010.

[17] J. Rao and K. A. Ross, "Making b+- trees cache conscious in main memory," in *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2000, pp. 475–486.

[18] S. Chen, P. B. Gibbons, and T. C. Mowry, "Improving index performance through prefetching," in *SIGMOD '01: Proceedings of the 2001 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2001, pp. 235–246.

[19] I.-h. Lee, J. Shim, S.-g. Lee, and J. Chun, "Cst-trees: cache sensitive t-trees," in *DASFAA'07: Proceedings of the 12th international conference on Database systems for advanced applications*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 398–409.

[20] L. Gruenwald and M. Eich, "Database partitioning techniques to support reload in a main memory database system: MARS," in *International Conference on Databases, Parallel Architectures and Their Applications,. PARBASE-90*, 1990, pp. 107–109.

[21] ——, "Choosing the best storage technique for a main memory database system," in *Information Technology, 1990.'Next Decade in Information Technology', Proceedings of the 5th Jerusalem Conference on (Cat. No. 90TH0326-9)*, 1990, pp. 1–10.

[22] S. Moldovan, "Databases: towards performance and scalability," 2008.

[23] A. Salminen and F. W. Tompa, "Requirements for xml document database systems," in *DocEng '01: Proceedings of the 2001 ACM Symposium on Document engineering*. New York, NY, USA: ACM, 2001, pp. 85–94.

[24] R. Bourret, "XML and Databases," 2003.

[25] Software AG., "Tamino XML database," http://www.softwareag.com/tamino/.

[26] W. Meier, "exist: An open source native xml database," in *Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web, Web-Services, and Database Systems*. London, UK: Springer-Verlag, 2003, pp. 169–183.

[27] H. V. Jagadish, S. Al-Khalifa, A. Chapman, L. V. S. Lakshmanan, A. Nierman, S. Paparizos, J. M. Patel, D. Srivastava, N. Wiwatwattana, Y. Wu, and C. Yu, "TIMBER: A native XML database," *The VLDB Journal*, vol. 11, no. 4, pp. 274–291, 2002.

[28] S. Pal, I. Cseri, O. Seeliger, G. Schaller, L. Giakoumakis, and V. Zolotov, "Indexing xml data stored in a relational database," in *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*. VLDB Endowment, 2004, pp. 1146–1157.

[29] P. O'Neil, E. O'Neil, S. Pal, I. Cseri, G. Schaller, and N. Westbury, "Ordpaths: insert-friendly xml node labels," in *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2004, pp. 903–908.

[30] D. Adams, "Oracle xml db developer's guide. 11g release 1 (11.1)," Oracle, Tech. Rep., 2008.

[31] I. Taranov, I. Shcheklein, A. Kalinin, L. Novak, S. Kuznetsov, R. Pastukhov, A. Boldakov, D. Turdakov, K. Antipin, A. Fomichev, P. Pleshachkov, P. Velikhov, N. Zavaritski, M. Grinev, M. Grineva, and D. Lizorkin, "Sedna: native xml database management system (internals overview)," in *SIGMOD '10: Proceedings of the 2010 international conference on Management of data*. New York, NY, USA: ACM, 2010, pp. 1037–1046.

[32] T. Milo and D. Suciu, "Index structures for path expressions," in *ICDT '99: Proceedings of the 7th International Conference on Database Theory*. London, UK: Springer-Verlag, 1999, pp. 277–295.

[33] A. Arion, A. Bonifati, I. Manolescu, and A. Pugliese, "Path summaries and path partitioning in modern XML databases," *World Wide Web*, vol. 11, no. 1, pp. 117–151, 2008.

[34] B. Cooper, N. Sample, M. J. Franklin, G. R. Hjaltason, and M. Shadmon, "A fast index for semistructured data," in *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 341–350.

[35] D. Knuth.

[36] A. Arion, V. Benzaken, I. Manolescu, and Y. Papakonstantinou, "Structured materialized views for xml queries," in *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007, pp. 87–98.

[37] N. Tang, J. X. Yu, H. Tang, M. T. Özsu, and P. Boncz, "Materialized view selection in xml databases," in *DASFAA '09: Proceedings of the 14th International Conference on Database Systems for Advanced Applications*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 616–630.

[38] A. Balmin, F. Özcan, K. S. Beyer, R. J. Cochrane, and H. Pirahesh, "A framework for using materialized xpath views in xml query processing," in *VLDB '04: Proceedings of the Thirtieth international conference on Very large data bases*. VLDB Endowment, 2004, pp. 60–71.

[39] K. Aouiche, P. Jouve, and J. Darmont, "Clustering-based materialized view selection in data warehouses," in *Advances in Databases and Information Systems*. Springer, 2006, pp. 81–95.

[40] N. Tang, J. X. Yu, M. T. Ozsu, B. Choi, and K.-F. Wong, "Multiple materialized view selection for xpath query rewriting," in *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 873–882.

[41] I. Tatarinov, S. D. Viglas, K. Beyer, J. Shanmugasundaram, E. Shekita, and C. Zhang, "Storing and querying ordered xml using a relational database system," in *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2002, pp. 204–215.

[42] M. Ramanath, J. Freire, J. Haritsa, and P. Roy, "Searching for efficient XML-to-relational mappings," *Database And Xml Technologies*, pp. 19–36, 2003.

[43] S. Amer-Yahia and M. Fernandez, "Overview of existing XML storage techniques," *submitted for publication*, 2002.

[44] W. Lian, D. W. lok Cheung, N. Mamoulis, and S.-M. Yiu, "An efficient and scalable algorithm for clustering xml documents by structure," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 82–96, 2004.

[45] R. Nayak, "Fast and effective clustering of xml data using structural information," *Knowl. Inf. Syst.*, vol. 14, no. 2, pp. 197–215, 2008.

[46] W. H. Inmon, *Building the Data Warehouse,3rd Edition*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[47] S. Chaudhuri and U. Dayal, "An overview of data warehousing and olap technology," *SIGMOD Rec.*, vol. 26, no. 1, pp. 65–74, 1997.

[48] A. Sen and A. P. Sinha, "A comparison of data warehousing methodologies," *Commun. ACM*, vol. 48, no. 3, pp. 79–84, 2005.

[49] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. Sen Sarma, R. Murthy, and H. Liu, "Data warehousing and analytics infrastructure at facebook," in *Proceedings of the 2010 international conference on Management of data*. ACM, 2010, pp. 1013–1020.

[50] A. Datta, D. VanderMeer, and K. Ramamritham, "Parallel star join + dataindexes: Efficient query processing in data warehouses and olap," *IEEE Trans. on Knowl. and Data Eng.*, vol. 14, no. 6, pp. 1299–1316, 2002.

[51] D. W. Cheung, B. Zhou, B. Kao, H. Kan, and S. D. Lee, "Towards the building of a dense-region-based olap system," *Data Knowl. Eng.*, vol. 36, no. 1, pp. 1–27, 2001.

[52] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, and B. Becker, *The Data Warehouse Lifecycle Toolkit*. Wiley Publishing, 2008.

[53] N. Karayannidis and T. Sellis, "Hierarchical clustering for olap: the cube file approach," *The VLDB Journal*, vol. 17, no. 4, pp. 621–655, 2008.

[54] P. O'Neil and D. Quass, "Improved query performance with variant indexes," in *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1997, pp. 38–49.

[55] A. Guttman, "R-trees: a dynamic index structure for spatial searching," in *SIGMOD '84: Proceedings of the 1984 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1984, pp. 47–57.

[56] J. T. Robinson, "The k-d-b-tree: a search structure for large multidimensional dynamic indexes," in *SIGMOD '81: Proceedings of the 1981 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1981, pp. 10–18.

[57] M. Freeston, "A general solution of the n-dimensional b-tree problem," in *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1995, pp. 80–91.

[58] R. Bayer, "The universal b-tree for multidimensional indexing: general concepts," in *WWCA '97: Proceedings of the International Conference on Worldwide Computing and Its Applications*. London, UK: Springer-Verlag, 1997, pp. 198–209.

[59] P. O'Neil and G. Graefe, "Multi-table joins through bitmapped join indices," *SIGMOD Rec.*, vol. 24, no. 3, pp. 8–11, 1995.

[60] R. Armstrong, "Data warehousing: Dealing with the growing pains," in *ICDE '97: Proceedings of the Thirteenth International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 1997, pp. 199–205.

[61] D. Theodoratos and M. Bouzeghoub, "A general framework for the view selection problem for data warehouse design and evolution," in *DOLAP '00: Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP*. New York, NY, USA: ACM, 2000, pp. 1–8.

[62] J. Wang, S. Wu, H. Gao, J. Li, and B. C. Ooi, "Indexing multi-dimensional data in a cloud system," in *SIGMOD '10: Proceedings of the 2010 international conference on Management of data*. New York, NY, USA: ACM, 2010, pp. 591–602.

[63] S. Eo, Y. Li, H. Kim, and H. Bae, "Two-Tier Storage DBMS for High-Performance Query Processing," 2008.