SPATIAL-TEMPORAL QUERY PROCESSING

REVISION 1.4

Introduction

WhamTech's External Index and Query (EIQ) Products™ use a unique combination of technologies to index data, and process queries, including table-joins across highly normalized databases and multiple databases. EIQ Product indexes are designed to be highly distributed and scale to billions of records and terabytes of data, INSERT/UPDATE/DELETE 100s to 10s of 000s of records per second in real-time, AND simultaneously allow complex queries by a large numbers of users. One of the many features of the index and query technologies is the ability to efficiently process range queries extremely fast, and therefore process spatial (or GIS/location) and temporal (or time-series) data very fast. EIQ Product indexes are offered in three versions:

- EIQ Server® for multiple data source information sharing system implementation,
- EIQ SuperAdapter™ as an alternative single data source adapter in an existing or other-vendor information sharing system, and
- EIQ TurboCharger™ to accelerate and otherwise improve query processing on a single data source

Index and Query Technologies

WhamTech's index and query technologies are based on a combination of the three "Bs" of proven computer technologies to provide extremely high performance at low total cost; a) balanced Binary tree indexes, b) Bitmap representations of record number pointers that match queries, called Collections, and c) Boolean combinations of integer list and actual bitmaps. Others have implemented these three technologies individually to varying degrees of success over the years, however, we are not aware of anyone who has solved the scalability issues with real-time updates on balanced binary trees or successfully implemented the dynamic combination of the three technologies, as WhamTech has. The way the three technologies work together is that the balanced binary trees are queried, the list information obtained from the trees is used to generate Integer or Bitmap Collections (see Figure 1 over the page), and the Collections are Boolean-combined to generate interim or final result-set Collections that are then used to read data from data sources.

Balanced binary trees are a technology from the 1960s and the attraction then, as now, is that binary searches are considered to be the fastest method of searching ordered lists[1]; however, they are usually dismissed in Database 101 due to three major problems that forced researchers to move on to other forms of index trees, such as B-trees and B+ trees:

- Levels tend to get very deep, whereby a billion nodes, for instance, need 30 levels; this translates into time to traverse
- Rebalancing and rotation after an insert or delete can take considerable time and a very large number of nodes can be affected
- A worst-case scenario of deletion of a top node

[1] C. William Gear, Applications and Algorithms in Computer Science, Science Research Associates, Inc, 1978, p. A107.

WhamTech predecessor companies solved all of the above problems in a unique and proprietary manner. The core tree code has been stable and untouched for over 10 years, and is used in a number of technology areas by WhamTech.

Spatial Queries

To cope with spatial queries, other relational database management systems (RDBMSs) use specialized third-party data cartridges (or blades/plug-ins), which are usually separate specialized databases with (usually one-to-one) connectors to records in the main database. For example, for GIS information, the ESRI data cartridge is very popular. Range queries on normal RDBMSs are usually very inefficient, are slow and resource-intensive, as they create large temporary tables.

EIQ Products work with spatial and time-series data like any other data type, which allows for all data to be consistent and contiguous across the same database. EIQ Products accomplish this by working with range data differently from other database technologies: a range query is a straightforward sequential read of a balanced binary tree index from the first to last record in the range, resulting in a list of only record numbers that match the query. This allows EIQ Products to process queries on spatial and time-series data very fast. See Figure 1 below:

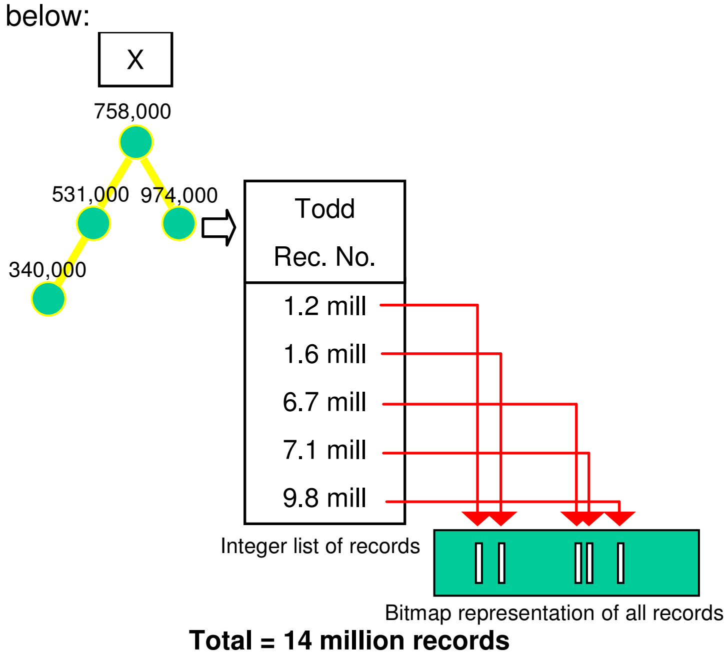

Figure 1: Balanced binary tree to Bitmap Collection

As a trivial example, as shown in Figure 1 above, a range query for X >= 400,000 AND X < 800,000, WhamTech would seek the first node that starts with 400,000, does not find one, but finds and reads the next higher node where X = 531,000, and traverses the tree to the node X = 758,000, reading and populating a Collection with record numbers as it goes from node to node. The node X = 974,000 would be checked but not read, as it would fail the X < 800,000 criterion.

A similar query would be performed for Y, and then the X and Y Collections and other Collections resulting from queries on other database data such as content, would be combined in a Boolean AND operation to arrive at a final result-set that can be used to read ONLY the data that is in the final result-set.

Spatial Query Examples

Given a specific location, X, Y, a certain type of service is sought within a certain area. These areas could be a square, rectangle, polygon, circle or any other shape that can be mathematically defined.

Example of a rectangle or square

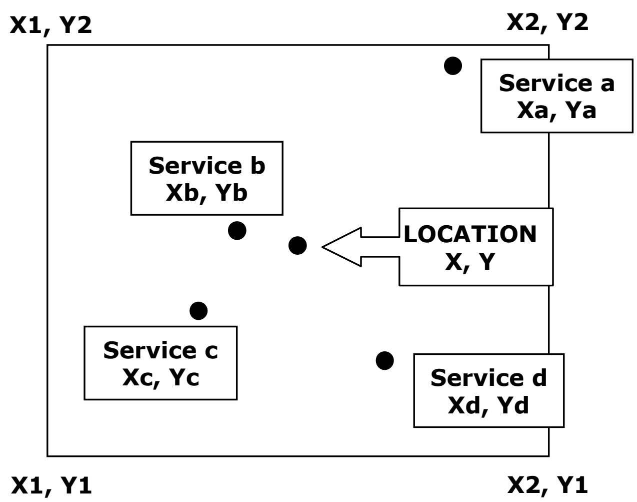

Figure 2: Query or search within a square

Assuming a rectangle or square is centered around a particular location X, Y, two range queries are made to find a certain type of service, where X1 ≤ X ≤ X2 and Y1 ≤ Y ≤ Y2. The created X and Y Collections along with Collections created with queries for the specific services sought, are combined with a Boolean AND to create a result-set Collection, which is then used to read ONLY the data for the services within the query rectangle or square; in this case, Services “a”, “b”, “c” and “d”. These services can be ranked according to the user’s criteria, e.g., distance from location X, Y, resulting in a listing of Services “b”, “c”, “d” and “a”.

Examples of non-square/rectangle or non-orthogonal areas with a small result-set

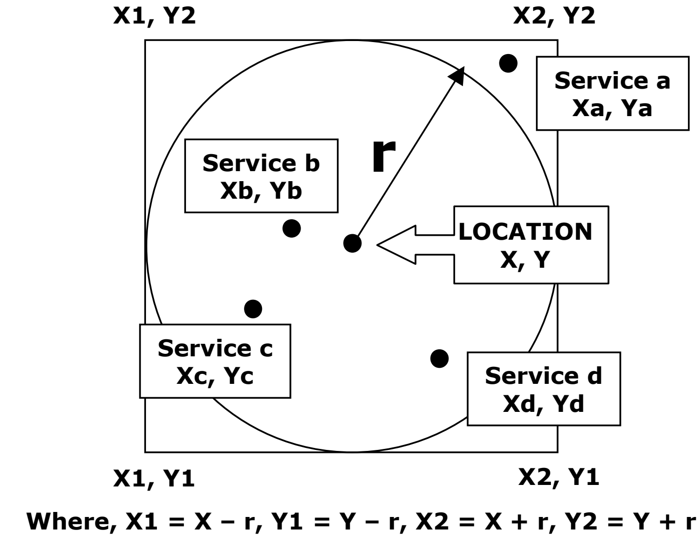

Figure 3: Query or search within a circle for small result-sets

In the example of a circle in Figure 3 above, typical of location-based queries, a square can be superimposed on the circle and queried in the same manner as the preceding rectangle or square example; however, result-set data that is outside the circle can be excluded from the ranked results if greater than distance, r, from the center location X, Y. The services can be ranked according to the user's criteria, e.g., distance from location X, Y, resulting in a listing of Services "b", "c", and "d"; Service "a" would be excluded.

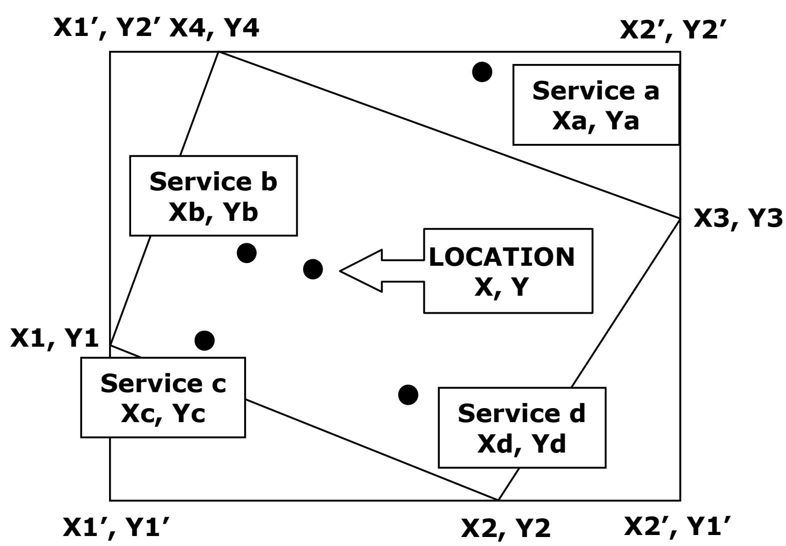

Figure 4: Query or search within a polygon for small result-sets

The polygon example in Figure 4 above, illustrates an approach similar to the circle example illustrated in Figure 3, whereby, result-set data can be excluded if it falls outside the polygon area. As long as the query or search area can be mathematically defined, result-set data can be excluded. This approach can be taken as along as the result-set data is small, as the computations needed to exclude data can cause overhead. For larger result-sets, an alternative approach may be needed.

Examples of areas with large result-sets

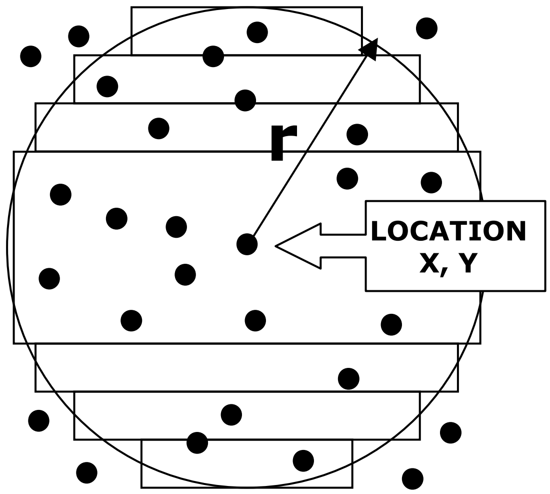

Figure 5: Query or search within a circle for large result-sets

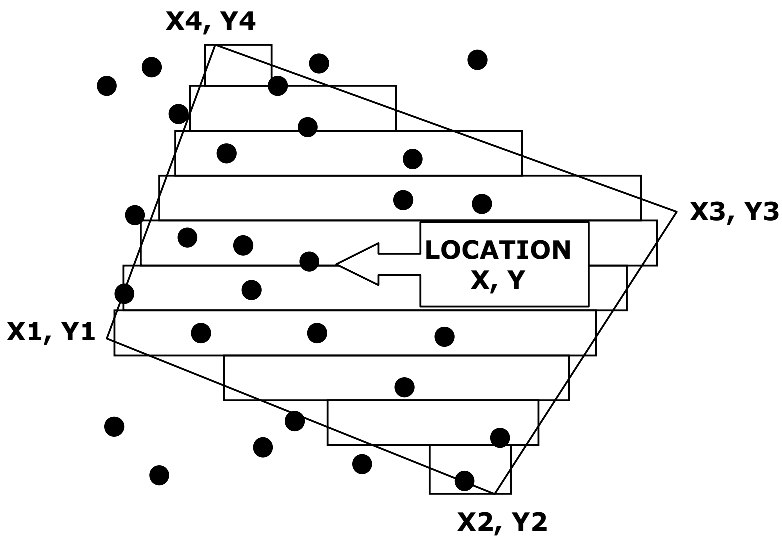

Figure 6: Query or search within a polygon for large result-sets

In some cases, if a large result-set is likely, the computation overhead to exclude data may be excessive, therefore a "cake layer" approach can be employed, as illustrated in Figures 5 on the preceding page and 6 above, whereby multiple rectangular queries are performed and then combined in a Boolean OR to only include result-set data – no computations are needed to exclude result-set data.

WhamTech Experience With Spatial Data

WhamTech was originally created as a company to develop an oil and gas application, WhamSDAT, which pre-processes 3D multi-component seismic. WhamSDAT is a great example of the benefits of WhamTech's spatial query processing, where a process was taking 24 hours on an IBM mainframe, took just over 2 hours on a 1996 PC with WhamTech.

Temporal (or Time-Series) Data

Temporal data occurs in almost all operational and transactional databases. Any record with a timestamp falls into this category. All records are temporal to some extent as they are created at a certain date and time, and are valid for a period until they are either updated or deleted, or superceded by a more up-to-date record. Two types of temporal data exist: transaction time, which is a data and timestamp, and valid time, which is the period a record is valid.

Problems arise when the database is queried for specific temporal data, e.g., hour, day, month, and year. Period data can be indexed and used to access temporal data; however, more complex temporal queries are normally required, particularly for data mining and business intelligence. Other database technologies overcome temporal data processing problems by creating a period table in a star schema or similar, which requires pre-processing and data warehousing, and then linking most other data to the period table through the fact table, as illustrated in Figure 7 on the next page.

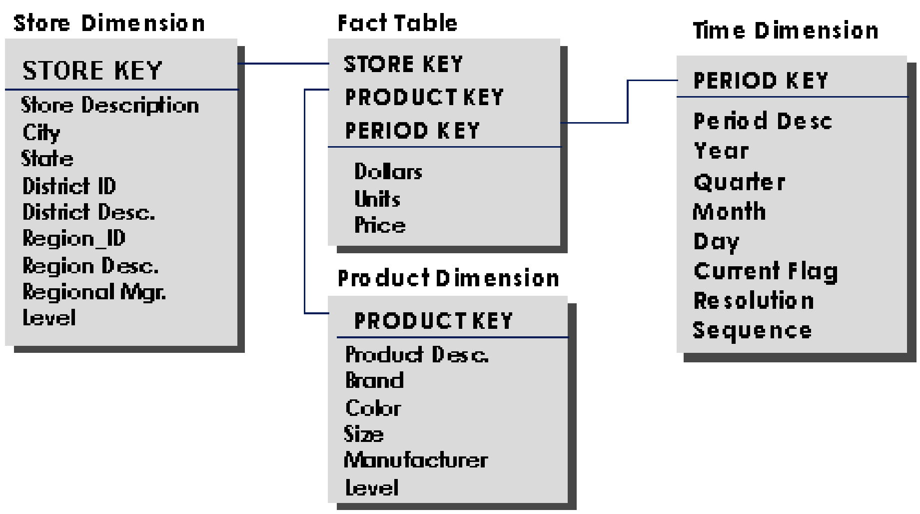

Figure 7: Classic database/ data warehouse/data mart Star schema

As WhamTech indexes columns individually, complex queries are resolved using Boolean operations across columns, in effect; creating composite indexes and table joins on the fly at the point of query. Plus, as discussed in the spatial section of this paper, a range query is efficiently executed as a straightforward sequential read of a balanced binary tree index from the first to last record in the range, resulting in a list of only record numbers that match the query – no interim or temporary tables. Similar to the spatial section of this paper, the result-set Collections can be many, and can be combined using Boolean operations to create a final result-set that is used to read only the final result-set records.

Very often, spatial and temporal data need to be combined together and with other data, e.g., “show me all Italian restaurants within five miles with a dinner-for-two special this week and rank by distance from my location”. Another example is a moving object or objects; space and time are involved. It is even possible to track moving objects in 3D by introducing a fourth dimension in addition to X, Y and time: Z.

Summary

WhamTech's EIQ Products use a unique combination of technologies to index data, and process queries, and can scale to billions of records and terabytes of data, with high rates of record INSERT/UPDATE/DELETE, AND simultaneously allow complex queries by a large numbers of users. One of the many features of the index and query technology is the ability to efficiently process range queries extremely fast, and therefore process spatial (or GIS/location) and temporal (or time-series) data very fast. Using relatively simple query methods, spatial and temporal data can be retrieved and ranked very quickly in almost all situations, including combinations of content, space and time.

For more information, please contact:

Gavin Robertson, CTO
WhamTech, Inc.
4450 Sojourn Dr., Suite 200
Addison, TX 75001
972-380-4645 x223
gavin.robertson@whamtech.com