

Technology Audit

Databases

CopperEye Greenwich

Written by: Michael Thompson

Date: January 2006

Abstract

Greenwich is a software solution for the storage, retrieval, and management of read-only data generated by business events. CopperEye uses the term 'business event data' to describe the type of data addressed by this solution, but might be better considered as 'event stream data'. Greenwich is a remote database server that provides two services against the underlying file systems. The first of these is an indexing service that discovers new data files, parses them, and then creates the required indexes which opens up the files for querying through SQL. The second service is the query service. This service takes standard SQL queries that can be generated from any source utilising ODBC, and optimises the query for best service using internal algorithms. In this manner, Greenwich reduces the cost of maintaining RDBMS for those types of data that do not require the full functionality of such systems. Data from structured event streams can now be stored in log files utilising inexpensive discs, yet data can be retrieved using SQL, as though the underlying file system was a relational source. Greenwich answers a clear business problem, in respect allowing organisations to fully comply with data storage and retrieval compliance issues, without the cost both in terms of hardware and maintenance typically associated with RDBMSs. Deployment of the solution is both quick and simple in general terms, but Butler Group believes that the solution has such an impact on the underlying storage system that additional benefit would be gained by undertaking a detailed evaluation of the storage subsystem requirements as a pre-deployment exercise. Although Greenwich is a new product, it utilises CopperEye's proven indexing technology that has been implemented in various guises over the past several years; including a DataBlade for Informix.

KEY FINDINGS

Key: ✓ Product Strength ✗ Product Weakness ⓘ Point of Information

✓ The utilised indexing technology is proven at the enterprise level.	✓ Provides SQL access to non-relational file systems.
✓ Allows organisations to fully comply with data storage and retrieval requirements.	ⓘ Organisations should take the opportunity to evaluate storage requirements.
✓ Deployment is quick and easy, with maintenance tasks highly automated.	✗ CopperEye's own SQL client only presents results in a scrolling list format.

LOOK AHEAD

Although the product is in itself fully featured, and allows SQL-based queries to be generated from other applications, there is an opening for CopperEye to provide additional search facilities. These will be available with CopperEye Search (due for release in Q1 2006).

► FUNCTIONALITY

Product Analysis

Greenwich is a software solution for the storage, retrieval, and management of read-only data generated by business events. CopperEye uses the term *'business event data'* to describe the type of data addressed by this solution, but might better be considered as *'event stream data'*.

In order to fully understand the power of Greenwich, it is necessary to understand the genuine business problem that it overcomes. Essentially, there are three distinct ways in which data can be stored: a data warehouse, relational, or log files. The organisational usage of each is decided upon by a combination of three factors: cost (hardware and management), usage requirement (the ability to search, modify, etc.), and the requirements imposed by legislation on the retention of data.

Transactional data has found a natural home in relational systems, data for BI purposes is most often found in data warehouses, while event stream data is most commonly stored within log files. Event stream data are those large data volumes that are generated during the normal course of business. A typical example would be the data generated by mobile phone companies in the course of their operation.

There are legislative needs for the storage and retrieval of all types of data, but when it comes to event stream data, the second half of the requirement is being ignored. This is due to the inability of log files to handle structured search. There have been many high-profile cases where the inability to retrieve required data in a timely fashion has led to the imposing of fines running into millions of dollars. It is becoming apparent that organisations are offsetting the risk of being *'found out'* and the subsequent financial impact against the cost of implementing storage infrastructures to handle the second half of the legislative requirements – that of search and retrieval.

There is a strong analogy here with financial institutions accepting the loss associated with credit card fraud, measuring it against the cost of implementing infrastructures that could reduce the loss caused by fraudulent transactions. Just as these institutions have finally had to succumb to pressure and implement the technology, so will other institutions have to stop paying *'lip service'* to the full intent of legislation and will have to implement storage infrastructures and solutions to address the complete issue.

It is not just a question of intent. There is a dichotomy in the technical infrastructure required for handling the terabytes of event stream data generated during normal operational processes. Log files are ideal due to their ability to handle large volumes of data, they are also ideal due to the fast loading and availability of data. The only area of storage requirements where they fall down is the ability to carry out structured queries against the data. This is where Greenwich comes into the picture; it fills in the missing piece of the puzzle.

Greenwich provides SQL queries against log file data that has been indexed using CopperEye's established core technology. Previous implementations of this core technology have been as a SDK, a solution for fast loading of data into Oracle databases, and as an Informix DataBlade. This proven technology is now available within a solution that addresses a wider audience, and a much more immediate and pervasive problem.

Product Operation

Greenwich is a remote database server that provides two services against the underlying file systems. The first of these is an indexing service that discovers new data files, parses them, and then creates the required indexes which opens up the files for querying through SQL.

The second service is the query service. This service takes standard SQL queries that can be generated from any source utilising ODBC, and optimises the query for best service using internal algorithms. Once the query has been optimised the underlying file structure is accessed (even though the underlying files might not inherently be ODBC accessible), the returned result set is formatted ready for presentation back through ODBC as a Table View. In this manner the underlying file structure can become SQL compliant, without the need to move or modify data in any way.

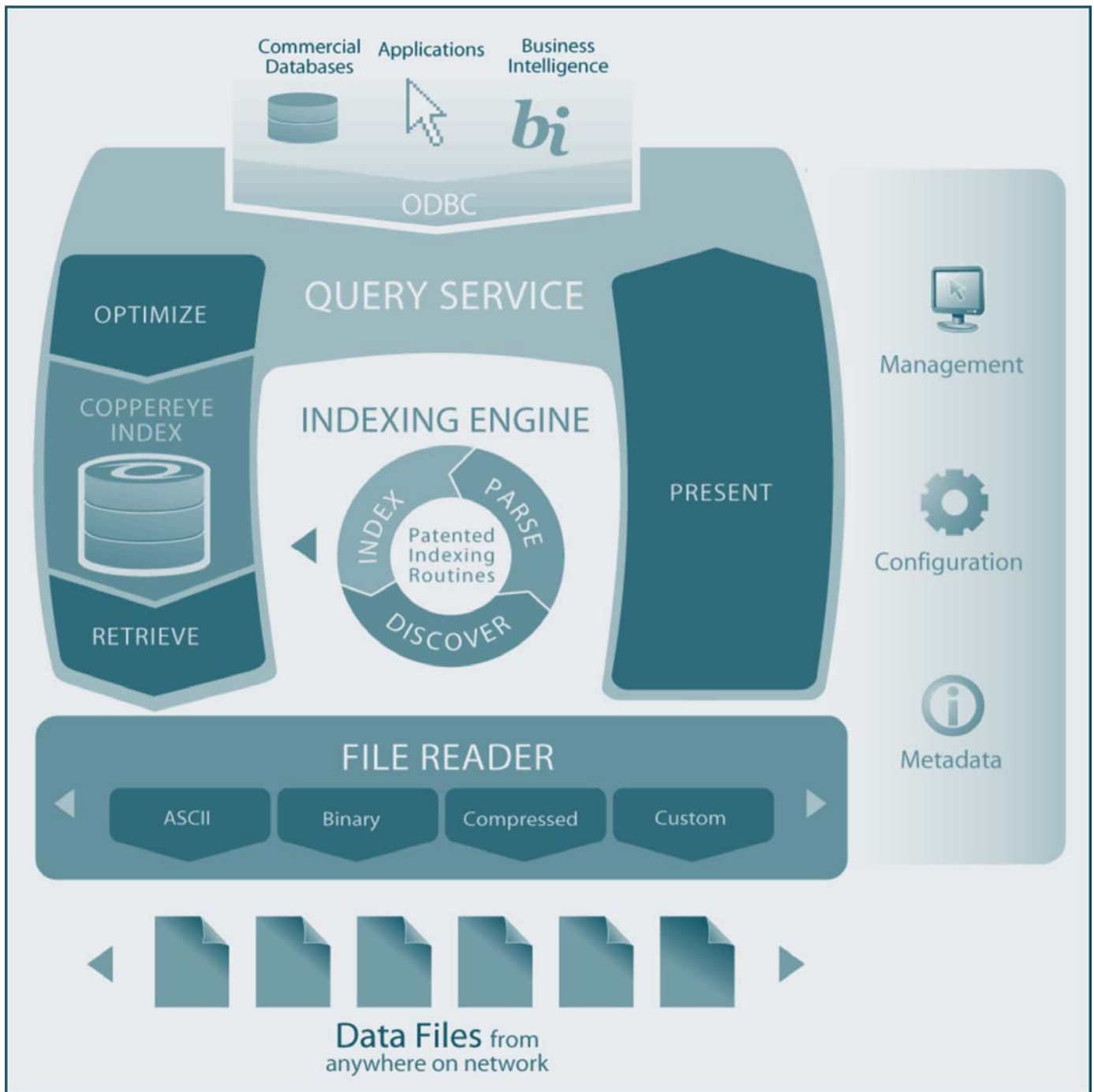


Figure 1 – Greenwich Architecture

There are aspects of the indexing and query services that require some further explanation. Each index created by Greenwich records the field value, the file identity, and the record offset. This allows the record to be retrieved in a single I/O operation when a SQL query is made that references a created index.

The indexes created are stored as binary files within the underlying file system, and these indexes utilise CopperEye's core technology that obviates the need to create expensive and complex disc subsystems to handle the overhead typically associated with index management in high throughput systems.

Within relational systems, there has always been a requirement to balance index creation against the overhead of INSERT when a large number of indexes existed. This becomes unmanageable when one starts to work with event stream data. If that data is indexed in the usual manner (for example, using B-Tree indexes), then the backlog of files waiting to be inserted would become unmanageable and the system would effectively never *'catch up'*.

CopperEye's indexing technology removes this overhead, which forms the ability for Greenwich to index log files that can then be accessed as though they were relational. In order for this to work, there has to be a defined or definable structure to the event stream (and by extension, the data carried by that stream). This is an important aspect of the usability of Greenwich. Although it can handle large volumes of data (and by large, we mean volumes that would bring any other structured file storage system to a grinding halt), there has to be this structure.

As Greenwich provides the functionality of being able to search and retrieve log file data by indexing the data stored within the file system, then there has to be a method of identifying those data elements that require indexing. This can only happen when there is the aforementioned underlying structure to the event stream. This means there is the requirement to identify and model the event stream prior to implementation.

As the validity of this solution, from a technical point of view, is totally reliant upon the ability of Greenwich to create scalable indexing without additional overhead, it is worth expanding on this.

A single CopperEye index file can scale to hundreds of millions of record entries without any major impact on performance. It also implements automated index file partitioning, which increases the scalability of the solution to the point where it can handle terabytes of data.

There is another advantage to the indexing algorithms and structures used by Greenwich, and that is in the discs used both for storage of the index files and the event stream data itself. Implementation of Greenwich allows the use of more inexpensive discs as it removes the overhead that was typically handled by the introduction of expensive faster discs. Further than that, it extends the possibility of Massive Array of Idle Discs (MAID) past the standard implementation of Write Once, Read Occasionally (WORO) for these storage infrastructures. The savings associated with MAID can now be extended to a Write Once, Read Many (WORM) methodology.

As the indexing technology used by CopperEye within Greenwich creates low latency between the creation of a record (an event happening within the stream) and the storage of that record, then the time to access the data created comes closer to real time. As Greenwich handles event stream data, which by its very nature is ongoing, there is a requirement to ensure that there is no locking against the file system for querying as data is being inserted. The closest analogy to this would be record level locking in a relational system, and Greenwich provides this facility, so the data can be searched as new records are inserted.

Mention has been made about the optimisation within Greenwich during query submission, and this is worth some further consideration. Greenwich selects the most appropriate index(es) to use to fulfil a query. This optimisation is cost-based against the lowest I/O impact measured against possible query paths. Again, this reduces the overhead on the system.

Within a relational system, this cost optimisation is typically based on statistics generated and stored as separate entities. The updating of these is handed off as a database management task, which is an additional overhead. Greenwich works differently. Each index within Greenwich dynamically and automatically maintains its own statistics internally. Therefore, cost optimisation is always up to date and accurate.

Product Emphasis

Greenwich is designed to handle event stream data and, as far as possible and as far as necessary, bring the benefits of the relational model to log file data, without negatively impacting performance. The solution is concerned with taking data that is unchanging (data that forms records rather than data that is part of transactional systems) and allows that data to be retrieved for a multiplicity of purposes. As the interface for both input and output is through ODBC, Greenwich forms a bridge between underlying file systems that have historically been hidden from, or unusable by, BI tools, applications, and end users.

► DEPLOYMENT

Greenwich can be deployed on Linux, or any tier-1 UNIX platform. It provides and utilises a full query compliant core-2 ODBC interface for connectivity to clients, and is certified to support access by Oracle, DB2, and IDS databases, along with Cognos and Business Objects BI tools.

On deployment, the target file directories are identified, and from that point the indexing service discovers new data files as they appear on the identified file systems. In order for Greenwich to parse these files and create the required indexes, the file structures have to be predefined. From that point Greenwich automates all aspects of the maintenance and use of the system. Therefore, deployment is both quick and simple.

One of the key aspects of Greenwich is the ability to use larger, cheaper discs within the file system sub-structure, without loss of performance. Therefore, it would be of benefit during the deployment phase to consider the whole hardware structure, as massive cost savings can be made in this area.

As an example of this, CopperEye carried out an implementation that demonstrated the savings to be made in the hardware area. The requirements for this were to show equal or improved performance with a reduction in hardware requirements against a commercial RDBMS.

Data storage requirements were 8.9TB of data. In order to service queries against this data volume, the RDBMS required 16 CPUs against the 4 required by Greenwich. Index storage on the RDBMS was 6.8TB (which is a clear indication in itself of the problem that Greenwich overcomes) and required 111 spindles, as opposed to the 5 spindles required by Greenwich.

The total infrastructure cost to maintain the RDBMS solution was in excess of US\$1.3 million, and the Greenwich hardware requirements came in at US\$347,000. With nearly a US\$1 million difference in hardware cost, Butler Group would consider this aspect of deployment a primary requirement.

Therefore, to gain the full cost benefit (as opposed to pure usage benefit), we believe that a typical deployment scenario would, and should, involve a complete strategic plan for the underlying data storage requirements.

► PRODUCT STRATEGY

The CopperEye indexing technology has been available in various forms for several years now, and has been proven in working environments. Greenwich can be considered as a natural home for this technology, as it provides a clearly defined solution for a clearly definable problem.

Greenwich is a solution that will be invaluable to those organisations that require storage and retrieval of data generated by structured event streams. As this is the case it naturally fits within large organisations, especially those that have these clearly definable event streams. A primary target at the moment is within the telecoms market, especially the mobile sector of this market.

The ability of Greenwich to work within this market is clearly demonstrated by a case study provided by CopperEye as a proof of concept.

The requirement was for a mobile operator with 12 million subscribers handling 200 million calls a day. The legal requirement is for the records of those calls to be maintained for a year and for access to a single subscriber's call records to be sub 5 seconds. Additionally, for various reasons, some legislative and others operational, there was an added requirement that the data generated by a call (which is part of a total storage requirement of 17.5TB of data) should be stored and made available within 20 minutes.

Clearly, this provides two effective routes to market for Greenwich. There is a demonstrable cost saving in hardware requirements allied to an ability to retrieve data from underlying file systems that have, historically, been difficult to search using standard SQL tools.

► COMPANY PROFILE

CopperEye is a privately owned company, employing 25 staff, headquartered in Box in the UK, with USA offices in Stamford, Connecticut and San Mateo, California. Formed in 2000 to promote new indexing technology, the company had early success offering their core technology as a software developer's kit for bespoke applications. With the release of Greenwich, CopperEye has recognised a gap in the market and has provided its 'first' product solution applicable to a wide market.

In general terms, CopperEye positions itself as a provider of enterprise search solutions that enables its customers to quickly retrieve exactly the records they need from months or years of history and billions of business transactions stored on low-cost file systems. CopperEye differentiates from other enterprise search solutions by specifically addressing the needs of business transaction data that would otherwise require a database solution to provide rapid retrieval of specific records. This is in contrast to how CopperEye characterises other vendors in the search market whose focus is on unstructured data such as Web pages and word processing documents and who return search results that require end users to manually evaluate each result to determine its relevance. CopperEye's principle value proposition is the promise of more powerful applications that have dramatically enhanced access to historical data, while simultaneously realising an order-of-magnitude reduction in data management infrastructure costs.

The CopperEye solution comes to market at a time of government legislation and dramatically increasing transaction volumes across nearly all industries. Sarbanes Oxley, the US Patriot Act, or the EU Directive on retention of data – all industries are facing increasing regulation. Examples of the growth of data are everywhere from wireless data networks and the introduction of third generation (3G) wireless technology to the emergence of Radio Frequency Identification (RFID) throughout the supply chain. In this climate CopperEye states its go-to-market strategy is to focus initially on the telecommunications and Internet service provider markets and then expand into other markets. Current customers include Kingston Communications, MessageLabs, and several other firms that are currently not disclosed.

The recent announcement on the agreement in the EU that telecommunication companies and ISPs will be required to store call information data for two years (or longer), leading to a massive increase in data retention, is precisely the market that CopperEye is addressing, and the Greenwich solution appears to have a timely release.

► SUMMARY

Greenwich takes the power of CopperEye's proven indexing technology and embeds it into a software solution that answers a genuine business requirement. Given the legislative requirements on organisations to both store and *retrieve* data generated across numerous sources, it is important for those organisations to implement solutions that *fully* answer compliance requirements in the most cost-effective manner. Greenwich handles the storage and retrieval of data generated by business events (event stream data) that typically has had to find a home in expensive relational systems (which allow data retrieval), or less expensive log files, where data retrieval becomes far more problematic.

Greenwich is not designed to replace RDBMSs or log files for all aspects of data storage, but to give organisations a viable alternative to both for specific requirements. These requirements are concerned with data that is generated by structured event streams, and which requires rapid storage allied to the ability to search and retrieve specific data records based upon the use of standardised SQL queries.

Over the past few years, Butler Group has constantly reviewed CopperEye's indexing technology, and has always promoted the value of the technology. However, one of the problems associated with indexing technology is that it is very difficult to 'sell' as a business solution. With Greenwich, CopperEye now has a marketable complete solution that should find favour with many large organisations, and should, in our opinion, at least be carefully considered.

Contact Details

CopperEye Ltd

Suite 47, Aztec Centre
Almondsbury
Bristol
BS32 4TD
Tel: +44 (0)1454 203610
Fax: +44 (0)1454 203330
E-mail: contact@coppereye.com
www.coppereye.com

CopperEye

263 Tresser Blvd.
9th Floor
Stamford, CT 06901
USA
Tel: +1 (203) 564 1997
Fax: +1 (203) 564 1402



Headquarters:

Europa House,
184 Ferensway,
Hull, East Yorkshire,
HU1 3UT, UK

Tel: +44 (0)1482 586149
Fax: +44 (0)1482 323577

Australian Sales Office:

Butler Direct Pty Ltd.,
Level 46, Citigroup Building,
2 Park Street, Sydney,

Tel: +61 (02) 8705 6960
Fax: +61 (02) 8705 6961

End-user Sales Office (USA):

Butler Group,
245 Fifth Avenue, 4th Floor,
New York, NY 10016,
NSW, 2000, Australia USA

Tel: +1 212 652 5302
Fax: +1 212 686 2626

Important Notice

This report contains data and information up-to-date and correct to the best of our knowledge at the time of preparation. The data and information comes from a variety of sources outside our direct control, therefore Butler Direct Limited cannot give any guarantees relating to the content of this report. Ultimate responsibility for all interpretations of, and use of, data, information and commentary in this report remains with you. Butler Direct Limited will not be liable for any interpretations or decisions made by you.

For more information on Butler Group's Subscription Services please contact one of the local offices above.