THE CLIPPER GROUP
Navigator™
Navigating Information Technology Horizons℠
Published Since 1993 Report #TCG2006072 August 16, 2006

SAND/DNA Gives Tools to Temper the Hugeness of Enterprise Data

Analyst: Anne MacFarland

Management Summary

Many enterprises are overwhelmed by their business enterprises and are sinking rather than swimming. Even within the realm of structured data, a wide variety of data assets has become broadly relevant to a variety of business processes. New waves of relevant information loom large every time enterprises seek to address a new market. In addition, the standard warehouse that aggregates business information and feeds it to reporting and business intelligence applications is foundering in these seas of information while attempting to meet increasingly impatient requirements for instantaneous production of real-time data. As companies more rapidly consolidate operations and evolve their business strategies, the *how-ya-doin'* of key performance indicators (KPIs) must evolve, along with the cubes that are prepared for analysis and the reports that feed the dashboards. And this is just the data side of the problem – there is a business side as well.

Government regulations demand that enterprises keep more data and keep it safe. Stockholders demand reliable, repeatable brilliance and timely evolution of strategy with no embarrassing mistakes. More people, and more kinds of people, are analyzing business data, both to give their company an edge, and to give themselves an edge in their company. Many of them have a far shorter decision window than that envisioned by the inventors of data warehouses. Most of them are seeking an innovation with which to do better than others and succeed in the business. This innovation is seldom found using the reports and data schemas of yesteryear. So, the enterprise is harder to satisfy, and its demands are growing even faster than the sea of enterprise information.

SAND Technology, a company based in Montreal, Canada, has a product family called *SAND/DNA (Dynamic Nearline Architecture)* that makes data quicker to manipulate and less expensive to store. The data it manages is the vast majority of structured enterprise data that documents the enterprise and forms its up-to-the-minute institutional memory. Because this information is not subject to update, segregation of it in a separate repository can improve enterprise data use – *as long as that segregation does not remove it from active use*. SAND's SAND/DNA products feature *tokenization*, which shrinks the space (and money) that must be dedicated to maintaining this data, and a *column* (versus row) *orientation*, which compresses data more efficiently, and lets the data services that organize structured data be done more efficiently. SAND/DNA lets data be queried in a compressed state, and cooperates actively with front-end databases. SAND/DNA gives a way of managing read-only business data that can optimize the whole data environment for the opportunistic and changing use that today's economy demands. For more details, please read on.

IN THIS ISSUE

The Clipper Group, Inc. - Technology Acquisition Consultants ♦ Strategic Advisors
888 Worcester Street ♦ Suite 140 ♦ Wellesley, Massachusetts 02482 ♦ U.S.A. ♦ 781-235-0085 ♦ 781-235-5454 FAX
Visit Clipper at www.clipper.com ♦ Send comments to editor@clipper.com

The State of Enterprise Data

While it is traditional to talk about the lifecycle of information, and how it is less frequently used over time (with exceptions), it is sometimes more useful to talk about the *character* of data. Detail data is usually transaction data.[1] It is captured by business processes, order entry systems, click-stream monitors, RFID sensors, shop floor equipment, and the like. Almost immediately, it gets attributes, perhaps only a time stamp and a table or file location. Much of detail data is only used in the aggregate.

Aggregated data is often time based (e.g., monthly, calendar-quarter, annual) or geographically amassed. Aggregate data can be represented in tables and spreadsheets or built into multi-dimensional cubes[2]. Aggregated data feeds the reports, KPIs, and dashboards that tell people how the business is doing. But, it is a combination of aggregate and detail data that helps people to do their jobs, and, more specifically, to do their jobs better.

Many methods and structures have evolved to organize electronic information in useful ways. None of them is simple, quick, or easy, particularly as the bulk of data becomes large. If the data has not been cleansed of input errors, and the inconsistencies checked, the variant data will not be recognized or used, and the view of the data will be skewed. If the context of the data being analyzed is not consistent, the analysis will be meaningless.

Data is not universally relevant. This is obviously true of detail data, but is also true of data aggregates, for data aggregates often have been filtered for a specific intended use. So data must be used not just as a value, but as a value aggregated by the attributes that govern how, where and when it can be used. These attributes, usually expressed as metadata, have become a routine part of structured data operations, now that importing and aggregation of data from multiple sources has become a part of most forms of data analysis.

Traditional standard reports are aggregates of data structures, often generated differently for different recipient roles. They are often limited in scope, and can be produced - even with the bulk of modern business information - relatively quickly. Ad hoc reporting is less efficient, for the data must be readied for analysis in the context of the end user who needs it.

This is where the bulk of today's large enterprise data repositories become a problem. Even with the best of relational clarity, great metadata, and consistent organizational semantics, like negotiating the streets of large cities like Boston at rush hour, *it simply takes too long to get analysis done.* The increasing demand for ad-hoc queries and new kinds of reporting only make the situation worse. Data processes must become far more efficient to support the kinds of use that enterprises require.

The SAND Technology Approach

Much of the business use of information is initiated by search and query. Therefore, while off-loading of data to a secondary repository is beneficial in many ways,[3] this offload must be fully indexed and able to be queried. Because of the burgeoning costs of data storage, and because the sluggish, rush-hour response is to be avoided, some form of compression is clearly a good idea. SAND Technology accomplishes both these requirements in an elegant fashion. It takes a column-oriented view of data, and then tokenizes data entries, generally shrinking them 90% (think of the storage savings), while attributing their tokens with the full table and column-based indexing that allows the data to be queried in a compressed state. Let's look at these two strategies, and then step back to see what SAND does with them.

The Benefits of Column-Oriented Thinking

SAND's column-oriented architecture is not a new idea, but it is a clever one. Think of it this way. In most languages, writing is horizontal. Perhaps because of this, many approaches to structured information take a horizontal, row-based approach. It turns out there is a lot to be gained by re-orienting your thinking.

- Columns tend to be a rich source of *characterization* attributes, just as table organizations are a rich source of *context* attributes. If the table tells you what the values pertain to, the columns tell you what attributes are being documented. Rows might tell you the customer name, which for business process analysis, is interesting but not as interesting as the customer's attributes and the service challenges those attributes cause.

[1] It can also be content, as in medical images, but this kind of detail is beyond the purview of the SAND solution.

[2] Cubes are aggregates of related data that may be manipulated (also known as *slicing and dicing*) to analyze a particular part of business or organizational operations.

[3] Data from multiple repositories can be collocated. Background data utility services can identify and promote use of a system of record (often called *Master Data Management*), and identify and reconcile inconsistencies (sometimes called *Entity Analytics).* Administrators can identify new schemas and cube views to satisfy emerging patterns of demand by users.

- Columns tend to be of a specified data type, making checks for data quality that are done by column far more efficient than checks done by the horizontality of a row.
- Also, column mentality is more easily responsive to schema changes, because these changes usually are reflected in columns. Some new attribute becomes relevant, forcing a new column. If the repository has a row-based orientation, it is beset by occasions of inconsistency. Vertical thinking recognizes a new column, and either applies it to a limited domain of rows, or adds null values to the entire table.
- Cubes and KPIs are basic conveniences of structured data analysis. Like most conveniences, they don't repurpose easily. A column mentality makes it easier to build new cubes and re-architect key performance indicators to reflect new organizational realities.
- As records are added, a row-orientation puts the whole record in a single partition. A column orientation spreads the load, minimizing the rebalancing of data. It also makes tuning the database more straightforward.
- In many table columns, the values are limited.[4] A column-based orientation fosters tokenization, discussed below.

Column-based thinking moves data processes a step away from optimizing the *find* process based on the paper traditions of a master attribute, to a *find* process based on the Boolean database traditions of search and query. A change to column-based thinking is one way to meet organizational demands for quick response, and to address the challenge that use of business information as a management tool, not just a process tool, has caused.

Benefits of Tokenization

Tokenization pervades our existence in more than the acronyms that let us express complex ideas in a few letters. Think of the pollen that lands on corn silk, transferring minute amounts of DNA that will produce the kernels of next year's corn crop. Like molecular DNA, SAND/DNA has all the information, in the form of XML metadata, to restore data to its full contextual wholeness.

The processes of searching, finding, and staging the data have everything to do with data's attributes, but very little, at least on the first pass, with the actual values contained in the data. These and other data routines are far more easily done, and more quickly done, if the data is small. Tokenization bundles the contents, but still exposes the handles by which data is used. This makes a great deal of sense in large data environments – if the token is attributed richly enough to be useful.

The trick is that the token must include all the necessary information to use the information, and to restore it to its full representation. SAND has a heritage in doing just this, and a wealth of patents to support its methodology. By taking a column approach, SAND makes the tokenization[5] approach extremely efficient as well as effective.

SAND Compacted Tables

With tokenization and a column orientation, SAND produces compacted tables, together with a separate repository of metadata and mapping information. The SAND repository can be the back-end of multiple databases. Because its tokens are small, their use can solve the problem of contention for resources in many cases. While SAND/DNA does not solve the need to reprocess data for different kinds of analyses, its existence as a separate data tier gives a separate venue in which much of this work can be done - without affecting the production database. For new analyses that span data from both tiers, SAND/DNA gives an opportunity for parallelism. The full indexing of the tables allows the repository to be used as a target for queries as soon as it is loaded. The repository allows the relevant information to be quickly identified, and the queries against the tables to be processed without expanding the data itself. The nature of the SAND architecture allows federation of the SAND/DNA environment with whatever high-performance database (even an in-memory database) that an enterprise wishes to use. It is not a replacement for other databases, or just a passive storage offload to a different storage environment, but an active tool to optimize the use of a very large whole.

SAND/DNA Access

SAND/DNA Access is the complement to any data warehouse. It provides read-only access with full versioning and stores the metadata in XML format with the compacted tables. It is fully

[4] The most obvious example is one in which the values are yes or no, but colors and even numerical values tend to be within a range.

[5] SAND's tokenization is not unlike the way Terracotta introduces resilience at the session level in J2EE environments. While these products address very different parts of the IT environment, there is a similar emphasis on using deftness and smallness to make stressed environments work better. For more information on Terracotta, see **The Clipper Group Navigator** dated July 21, 2006, entitled *The Scale-Out Cornucopia – Terracotta Enhances IT Strategies of Plenty*, and available at http://www.clipper.com/research/TCG2006063R.

searchable with standard SQL tools. Its loading performance can exceed five TB per hour. Query performance exceeds 75 million rows per second. And the data itself shrinks to about 10% of its original size. Such are the virtues of compression and tokenization.

SAND/DNA Analytics

SAND/DNA Analytics is a data mart instantiation of SAND's *Nucleus* database technology. In today's environment, it finds its optimal use in coping with the ad-hoc queries and other forms of unanticipated analysis that beset the modern enterprise. SAND/DNA Analytics data can be used by popular business intelligence tools. It is easily updatable, with instant rollback and point-in-time consistency that SAND calls *Time Travel*. SAND/DNA Analytics has good performance against wide tables.[6] Because the data in SAND's environment is inherently indexed and because the data structures can support the addition of new columns without the need for wide-scale reorganization of existing data structures, SAND/DNA can help an enterprise's structured data environment accommodate new kinds of use.

Use Case Scenarios

SAND and CRM (or MDM)

Customer Relationship Management (CRM) is inherently grounded in detail data. Master Data Management (MDM) is an inherently aggregate strategy, where one version of the truth can be precipitated to many situations. The single version of the truth and the single view of the customer that underlie Master Data Management initiatives and Customer Relationship Management suites are both made more complex by the existence of multiple or rigidly partitioned databases. SAND/DNA maintains bidirectional data flows with the operational database. The metadata repository links the online and archived data so the two can be used as one.

In both CRM and MDM, there is a need to move easily from aggregate to detail data and back again in order to analyze a situation or opportunity. The ability to navigate at speed is well supported by the SAND approach.

In the *SAND/DNA aCRM* product, customer data is stored as one very wide table, including response to campaigns and other detail data. Because SAND/DNA Analytics has an analytic engine, there is no need to export data for analysis. In addition, SAND/DNA aCRM can use any tools that conform to the CRISP-DM[7] standard. The efficiency of SAND's SAND/DNA product can cut modeling time and make data exploration easier.

SAND and SAP

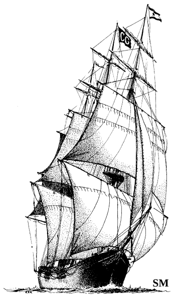
SAP is all about the business details – the assets, the processes, and the actual experience – and how to leverage that detailed knowledge to run the business well. This is another situation, like CRM, where the ability to navigate aggregations of information and layers or detail is very important. SAND has worked with SAP to optimize use of SAND/DNA as an offload environment for SAP BI. Because of the granularity and completeness of SAND's tokenized data, it is a very good pairing with SAP, with SAND providing a separate data layer where data can be stored in a consistent, granular, and historic form. SAND/DNA Access can be used directly by SAP BI tools as a data source. SAP's *Analysis Process Designer* can be used to build new data layers and new *InfoCubes* and *ODS Objects*. SAND has worked actively with SAP, its partners, and large system integrators to optimize the utility of SAND/DNA in SAP environments.

SAND and Data Bases

SAND/DNA, when used with traditional database management systems such as Microsoft, IBM DB2, Oracle, and NCR, lets administrators offload aging data, detail data, and/or any data that is more clutter than useful. It allows offload to tape. It easily accepts new data elements. Access restrictions are maintained in the SAND environment and a view can be created based on a union of tables from both environments.

Conclusion

The rising tide of demand for ad-hoc reporting, together with the size of large enterprise databases, is a no-win situation. Offloading read-only data, without a bifurcation of the search and query functionality, can provide the kind of response times that allow a business to survive and thrive amidst cut-throat competition. SAND/DNA provides a way to keep a very large volume of structured data searchable. If your rising tide of enterprise data is confounding your best efforts to use it, think columns. Think tokens. Think SAND/DNA.



[6] Enterprise tables are generally wide and getting wider (and, of course, SAND's ability to add columns fosters that trend). This is in contrast to the triple-store three-column databases that underlie the incredible performance of Web Search. The focus of Web Search is to *not* to manipulate.

[7] Cross Industry Standard Process for Data Mining

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Data Strategies and Information Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial "128" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** ***Captain's Log***, **The Clipper Group Voyager**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.