Sybase IQ Text Technologies: *A Technical Overview*

Seth Grimes

Alta Plana

Sponsored by

SYBASE®
An SAP Company

Updated March 2012

Text Technologies for the Sybase IQ Analytical Database System

Sybase IQ is the first leading analytical database management engine to embed text analytics functions within the database system, using ground-breaking information-extraction technology to power enterprise-scale analysis.

Text technologies

Sybase IQ has long been able to store large text files in binary and character large-object fields. While other DBMSes share this capability, Sybase IQ is unique among enterprise-grade, analytically focused systems in its provision of text analytics functions. These functions index and search text fields and facilitate integrated analysis of database-stored text and data. They position users to exploit embedded advanced analytical functions to support data mining of text sourced information and lay the groundwork for enhancements planned for subsequent Sybase IQ releases.

Further, Sybase provides options for Web-data acquisition and cleansing and for handling of text in a wide variety of binary formats. In data warehousing and analysis of numerical data, sourced from enterprise operational and transactional systems, roughly 80% of effort will typically go into data preparation: data acquisition, cleansing, and structuring for analysis. Similar figures apply in work with text-sourced data. Availability of a comprehensive technology set – text-data acquisition, processing, storage, access, and interactive analysis – speeds time to business insights whether obtained via data mining, BI reports and dashboards, or search.

Unstructured sources

Business intelligence has long focused on numerical data despite the very significant business value locked in text sources. Text captures customer, market, and stakeholder attitudes and opinions, the *Voice of the Customer* (and of the *Patient*, *Voter*, *Employee*, and *Market*). These voices, expressed in *natural language*, have until recently resisted automated analysis. With new, powerful text analytics capabilities, the analytics balance point has shifted toward the unstructured end of the information spectrum.

Businesses seek to extract insights from unstructured sources that may include:

- E-mail and text messages;
- Web pages, blogs, forums, and other social and news media;
- Contact-center notes, survey responses; warranty and insurance claims;
- Corporate reports and filings; and
- Legal documents and scientific literature.

Information may originate in enterprise CRM, content-management, e-mail, messaging, and file systems, on the Web, and on social platforms. Some are under enterprise control and others are not.

Further, applications are diverse, responding to a broad set of business needs. They range from intelligence and counter-terrorism; via risk and fraud management and e-discovery; to customer support, marketing, and brand and reputation management.

Text analytics processes

There are three basic steps in any application of text analytics:

1. Text acquisition.
2. Preparation, processing, and storage.

and

3. Analysis and presentation, whether involving:
   - Data mining,

CC BY ND

- Interactive data exploration,
- Business processes such as e-discovery, or
- Search

Given that text-sourced information complements enterprise transactional and operational data, the analysis and presentation step may include a significant data-integration effort and unified analysis of text and data: unified search, data mining, and BI.

Sybase IQ, text, and the analytics promise

There are many approaches to text analytics, designed for different business domains, goals, and users. Our interest, in this paper, is the integration of text sources into an enterprise analytics program to support analytics-reliant business processes, an area where Sybase IQ excels.

This paper elaborates on business benefits and technical challenges related to taming text, but its real focus is the technical implementation of text analytics and related processes on the Sybase IQ analytics DBMS platform. It covers, in particular:

- Text acquisition and processing.
- Text analytics functions embedded in Sybase IQ or provided via SAP BusinessObjects Data Services.
- Unified Sybase IQ management, search, and query of text and data.
- Interactive data exploration via visual interfaces.

It concludes with application scenarios: Business benefits of storing text in Sybase IQ.

CC BY ND

Why and When to Store Text in the Database

Information in text

A large proportion of business-relevant information – text, video, images, and audio – resides in binary formats. Because sources are "unstructured," their data content has until recently been difficult to extract and analyze. The situation has changed. Given the maturation of content technologies, enterprise users are now able to automate analytical processes that:

- **Generate metadata that identifies that content.** Metadata may include author, title, topic, keywords, file format, production/modification date, language, and a variety of other values.
- **Extract information from the content.** Focusing on text, information of interest, found within text documents, may include:
  - Named entities such as persons, organizations, geographic locations, and products.
  - Pattern-indicated entities such as e-mail and postal addresses, dates, telephone numbers, Social Security numbers, and currencies and quantities.
  - Complex features linking entities such as events, relationships, and sizes or measured amounts.
  - Structured data from data tables embedded in textual sources.
  - Topics and themes: What a given piece of content is *about*.
  - Sentiment – attitudes, emotions, opinions – associated potentially with any of the features given above.
- **Apply analytical techniques to create business insights** from extracted information.
- **Facilitate unified information access**, single queries, searches, and analyses that draw simultaneously on both textual information and conventional data tables.
- **Support interactive, visual, exploratory data analyses.**

Text and data for richer analyses

Text-sourced information complements data from transactional and operational systems. Transactional and operational data will tell businesses:

- ***What***: Product and service information, prices, costs, and inventory.
- ***Who***: Customer, supplier, retailer, and channel profiles.
- ***How*** and ***When***: Orders and sales, payments, telephone calls, hotel and air reservations, and Web logs.

This data tracks behaviors and is the raw material, whether collected and restructured in a data warehouse or analyzed in-place, for analytics – for business intelligence and data mining – that models and derives insights from the data. Transactional and operational data, however, lacks direct explanatory power. It does not tell you:

- ***Why***: Root causes, what motivated particular transactions and broader behaviors.

Text analytics is the key to this deeper level of understanding applying the steps above that discover and exploit ***information in text***.

Business benefits are maximized when you bring into play multiple information sources, applying appropriate analytical techniques to each source and to the ensemble. ***Unified analyses of text and data are richer analyses.*** They create ***lift***, insight that goes beyond what can be discovered with single sources and single techniques.

Enabling unified search and analysis of text and data

A variety of approaches will allow you to unify analyses of text and data. They

CC BY ND

include "federating" queries to repositories and databases that separately store text and data and joining the results in the application. For instance, you could leave text in a content-management system (CMS) or e-mail server or on the file server, in which case to join text and data, you would need to retrieve documents or records from the text repository, accessed via some query language, and data from the database via SQL, and then discover "key fields," perform the join, and filter results, in the client application. This approach does not offer the level of performance, nor the direct power and simplicity, of unified text-data storage within a single database where both text and data can be queried, searched, and analyzed with a single set of tools and with a single query language such as SQL.

Most modern database management systems, particularly enterprise-grade DBMSes, store both *fielded* data and unstructured text, the latter as binary or character large objects (BLOBs and CLOBs). Other field types such as CHAR, VARCHAR, BINARY, and VARBINARY store shorter character or binary, unstructured, values.

Large object fields (LOBs) may be retrieved but they are typically not directly queryable. They can typically be searched only by retrieving each row and scanning the fields' content at query time with an SQL function such as LIKE. Further, their information content, the ***information in text*** described above, can be analyzed only by retrieving the field contents to front-end applications via queries that utilize often-sparse and limited metadata. This process is slow and laborious and is essentially equivalent to a federated query approach.

Use of a text-optimized analytical DBMS enables far more efficient and powerful unified text-data query, search, and analysis than are possible with either a conventional DBMS (designed for transaction processing rather than data analysis) or an analytical DBMS that does not support text analytics.

The bottom line

Store text in the database when you need to query, search, or analyze both text and structured data tables. Use an analytical DBMS when your work is computationally intensive, involving big data and a significant calculation load.

CC BY ND

Sybase IQ text analytics architecture

A columnar, parallel "Big Data" analytical platform

Sybase IQ is a columnar, parallel data management platform designed for enterprise data volumes, user communities, workloads, and manageability.

- With a ***columnar DBMS***, physical data storage is optimized for queries that touch many values of relatively few database fields. The columnar approach optimizes storage and performance, in part through aggressive data compression, for data warehouses that support analytical workloads. Contrast the approach with row-oriented storage in a mainstream DBMS that was designed for transaction processing, which involves retrieving, updating, and inserting single rows in multiple data tables.
- Sybase IQ ***parallelization*** means both that multiple queries streams are handled simultaneously and that a given query may be split into threads that are processed in parallel – *many hands makes for quick work* – which is especially beneficial for operations conducted on many (sets of) data values.

  Sybase IQ parallelization is accomplished through a hybrid SMP-MPP architecture: symmetric multi-processing (SMP) that clusters multiple shared-memory computers and grid-based massively parallel processing (MPP) at the compute-node level.
- ***Enterprise user communities*** tend to be large and diverse with users representing different departments and job functions, carrying out a disparate set of tasks. An enterprise-class DBMS will deliver the interfaces, performance, and security required by enterprise user communities.
- ***Enterprise workloads*** are varied, including for analytics applications. They may include long-running data-mining and modeling tasks involving complex calculations, and they may include a large volume of relatively simple tasks that must be carried out in *real time*, for instance to support operational decision making.

The advantages of Sybase IQ's columnar, analytics-optimized DBMS architecture extend to work with text.

Sybase IQ text management

> *"The Unstructured Data Analytics Option extends the capabilities of Sybase IQ to allow storage, retrieval, and full text searching of binary large objects (BLOBs) and character large objects (CLOBs) within the Sybase IQ database."*[1]

For Sybase IQ, text fields are just fields, managed by the DBMS rather than by a content management system (CMS), e-mail server, or the file system. They do not break the relational model, which organizes data, from a logical point of view (from the database administrator and the application programmer sees the database), as data tables (relations) that may be joined according to certain key fields. The LONG BINARY and LONG VARCHAR types, which may extend up to 2 petabytes in size, conform to the Core level of the ISO/ANSI SQL standard.

Sybase IQ is an analytics engine whose columnar architecture answers the *Big Data* performance and scalability edge enjoyed by *NoSQL* database systems but with important advantages over *NoSQL key-value* stores:

- Text fields in Sybase IQ are compressed, which minimizes storage and data

[1] Unstructured Data Analytics in Sybase IQ, *http://infocenter.sybase.com/help/topic/com.sybase.infocenter.dc01268.1520/pdf/iquda.pdf*

CC BY ND

movement, boosting DBMS efficiency and performance and decreasing hardware requirements. The compression algorithm is appropriate for the text field type.
- The content of text fields may be queried, searched, and analyzed.
- Because Sybase IQ (also) stores data and metadata in structured, fielded form alongside text, it supports unified text-data query, search, and analysis.
- Sybase IQ embeds BI and advanced analytics functions within a parallel processing framework, under the control of a mature workload-management system, providing a high-level of analytics performance.

Sybase IQ facilities for fast, parallel, bulk loading of text (as well as structured data) provides an additional edge for IQ over other systems that manage text.

TEXT indexes

Sybase IQ supports full-text search via a TEXT index that stores positional information for terms in the indexed column. TEXT indexes may be created for columns of type CHAR, VARCHAR, and LONG VARCHAR, as well as BINARY, VARBINARY, and LONG BINARY, with support for UTF-8 UNICODE character sets (except for fields of CHAR and VARCHAR type). An optional text-configuration object regulates index-related settings.

TEXT indexes and the configurations that govern them are created and maintained via familiar SQL CREATE/ALTER/DROP. Configurations specify:

- STOPLIST, particular words or terms to ignore when indexing, typically common words such as "and" and "or".
- Minimum and maximum lengths of terms to be included in an index.
- A TERM BREAKER algorithm to use for separating column values into terms. The NGRAM alternative specified that n-character substrings be taken as terms for indexing; it is a choice that provides for the use of fuzzy matching in searches and queries.
- A *prefilter library*, which removes unnecessary information, such as formatting and images, from the document.

Sybase IQ additionally supports WORD (WD) indexes on CHAR, VARCHAR, and LONG VARCHAR (CLOB) columns. WORD indexes were designed to accelerate keyword searches with LIKE and CONTAINS clauses.[2]

File parsing with ISYS

Sybase partners with enterprise-search provider ISYS to bring the power of ISYS Document Filters, a set of connectors for extracting and rendering text from a wide variety of file, container, and email formats, to Sybase IQ users. The ISYS software is implemented as a *text prefilter* for the Unstructured Data Analytics option. Once configured, they operate automatically as part of a CREATE TEXT INDEX operation to parse character and binary text files, extract metadata and other text-captured features, and populate the IQ TEXT index.

ISYS Document Filters support every major, enterprise-relevant file format. They achieve processing speeds up to 40 GB per hour using a single processor and process both text and metadata. The Document Filters are suitable for a broad set of business domains with particular strengths in work with corporate-relevant sources and for analytics applications that include e-discovery, fraud detection, and forensic analysis.

Web extraction with Kapow

Sybase partner Kapow Software, whose Katalyst enterprise data integration platform

[2] Word (WD) Index in Sybase IQ, *http://www.sybase.com/detail?id=1057603*

CC BY ND

extracts data from Web, enterprise, and cloud sources and transforms the data for Sybase IQ loading.[3] Katalyst does not require that the source system provide an application programming interface (API). It does provide workflow and sophisticated transformation capabilities.

Web-data acquisition is not a simple task. The vast majority of Web pages were designed for human consumption, including "deep Web" pages generated from database-stored data via templates. They rely on HTML mark-up, and often on more exotic technologies such as Adobe Flash, that specify how page items should look, not what they are. Often pages of greatest interest can be reached only via menus and forms. Further, Web pages often contain extraneous material such as ads, with content embedded in a window or frame within a page. Nonetheless, handled properly, the Web and internal systems alike are extremely valuable sources of customer, supplier, product, competitor, and market data as well as of news, financial, and social information.

Katalyst works with sources as-published, even if designed for human rather than automated access. The Design Studio provides a visual user interface that lets you designate the portion of a Web page that contains data fields of interest. These fields aren't limited to explicit, labelled values. They may involve patterns and expressions, and the user can specify conversions. Further, Kapow can navigate menus and forms to reach pages of interest.

Kapow Katalyst then automates Web-data extraction and transformation, directly loading extracted information to Sybase IQ database tables. Kapow will populate both structured fields – values for date, title, author, and so on – and unstructured, free-text fields. This data, including the text fields, may be indexed for unified Sybase IQ text-data query, search, and analysis. It may also be further transformed post-loading – a data-acquisition approach known as *extract, load, transform* (ELT) – to facilitate analytical processing plans.

SAP Text Data Processing

Sybase IQ can take advantage of SAP's path-breaking Text Data Processing technology, based on Inxight text-analytics software first developed at the Palo Alto Research Center (PARC), part of the BusinessObjects solution set since 2007.

Text Data Processing (TDP) is part of SAP BusinessObjects Data Services. The software excels at information extraction from text. TDP makes sense of text by applying linguistic algorithms that decode word forms, parts of speech, and grammatical patterns. TDP also exploits sophisticated pattern-matching and statistical techniques. The result is strong ability to identify entities, facts, sentiment, and other *features*, as described in the *information in text* section of this paper. TDP information extraction turns text into data that can be stored and further analyzed in the Sybase IQ database. Other capabilities include automated summarization and document classification by use of taxonomies.

Sybase IQ search and query

Once data (fielded and full text) is loaded to a Sybase IQ database and indexed, it becomes fully searchable and queryable.

For text, Sybase IQ supports use of common string manipulation, substring, pattern-match, and length functions, including for large-object (LOB) columns.

The CONTAINS clause, used in the FROM or WHERE portion of an SQL query, provides a mechanism for full-text search, of terms, prefixes (partial terms), and

[3] Kapow Technologies automating job data collection, *http://youtube.com/watch?v=Uwk5ehcE2uw*

CC BY ND

phrases. CONTAINS supports:

- Simple expressions and query strings.
- Boolean conditions composed with AND, OR, and NOT operators.
- BEFORE and NEAR qualifiers that require one term to be within a minimum or maximum distance of another term.
- FUZZY matching, finding terms similar to those specified, for NGRAM TEXT indexes.
- Prefix terms, where a match is performed for a portion of an indexed term with a wildcard applied to the remainder of the term.

A key Unstructured Data Analytics advantage is that search terms and phrases are retrieved without scanning the data rows. Given that the TEXT index is used for term and phrase search, the user benefits from significant performance gains.

For retrieval of portions of text fields, functions such as LOCATE and CHARINDEX (location of string occurrence, with and without an offset) and PATINDEX (pattern location, with an optional offset), as well as substring functions, are particularly useful.

Further, a variety of stored procedures enable useful functions including chunking-out terms from text strings and generating statistics, including term frequencies, on TEXT indexes.

Sybase IQ analytics

Sybase IQ is a leading Big Data analytics engine with support for essential business intelligence, statistical, and data mining and modelling functions. The Sybase IQ-platform capabilities extend to analysis and presentation of text extracted information, on its own and in conjunction with analysis of data from structured sources. Sybase IQ runs, in-database, advanced analytics and data mining routines from Fuzzy Logix[4] and other partners. In-database execution is a high-performance, efficient, secure alternative to exporting data to external programs.

These in-database capabilities complement the use of SAP BusinessObjects Data Services – Text Data Processing and other analytical functions – providing users a set of options that meet a broad range of needs.

Modeling and model scoring operations can easily extend to cover text-extracted data – whether processed via SAP BusinessObjects Data Services or by routines embedded in the Sybase IQ engine – in addition to data from usual, structured sources.

[4] *http://m.sybase.com/files/Data_Sheets/SybaseIQ_FuzzyLogixInDatabaseAnalytics_ds.pdf*

CC BY ND

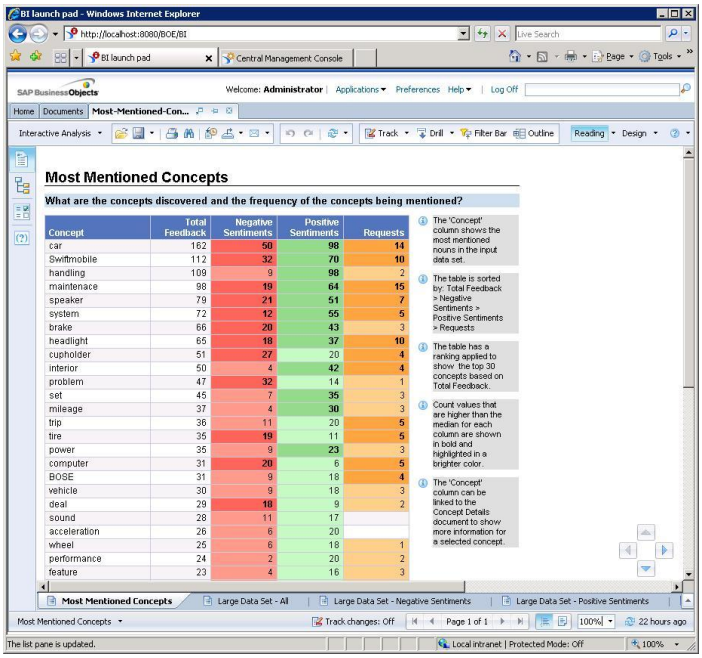
Search/Unified Analysis Scenarios

A number of usage scenarios illustrate possibilities enabled by Sybase IQ Unstructured Data Analytics (UDA) that can benefit enterprise users.

Voice of the Customer

Transactional and operational data – order, payment, and service records, business inquiries, customer profiles, financial transactions, call-detail records, Web-server logs, and the like – tell you the *What* and the *Who* of business interactions. To get to the *Why* – to understand satisfaction, motives, reactions, and plans – you need a different type of data. You need to get at information-rich *Voice of the Customer* sources.

Text analytics helps you extract and use Voice of the Customer (and of the Consumer, Market, Patient, Voter, and Employee) insights from the broad set of online, social, and enterprise unstructured data sources. These include online news, blogs, and reviews; social exchanges and status updates; and surveys, e-mail, instant messages, and contact-center notes. Sentiment – positive and negative feelings and emotions attached to *features* of interest (as discussed in the Information in Text section of this paper) – is a particularly important contributor, identified and computed via text analytics, in VOC analyses.



SAP BusinessObjects Data Services software provides specialized Voice of the Customer extraction capabilities. And the standard SAP BusinessObjects BI interfaces, backed by a Sybase IQ analytical database, provide an excellent environment for Voice of the Customer analyses that link text-derived entities, relationships, sentiment, and events, all insight elements, with data from transactional and operational systems.

Customer satisfaction and churn reduction

Process sales or customer-support e-mail messages, using UDA search and substring operations, to capture header fields (From, To, Subject, Date) as database fields and to extract product information from the e-mail body, the latter via search within messages for database-stored product names or SKUs.

Unified text-data analytics would allow association of e-mail inquiries, service requests, and other correspondence with customer profiles. This association would support sales/service representatives' proactively searching messages from customers whom predictive models identify as high-value, a cancelation risk, etc., to determine root causes behind transactions and to discern customer key interests. Imagine using Sybase IQ time series analytics to study trends in the volume over time of e-mail

CC BY ND

inquiries or questions about products or services, broken out by geography, customer type, sales channel, and other categories.

Regulatory and corporate compliance

Regulatory and corporate compliance typically entails ensuring that employees do not improperly disclose or act on confidential or proprietary information, which may include financials and news about pending deals, plans, and personnel movements. Often evidence of compliance violations is suggested by anomalous behaviors. Sybase IQ is an ideal platform for detecting suggestive patterns, and for investigating suspected infractions, via in-database analytics.

With UDA, assurance procedures can be extended to search of database-captured e-mail, messages, syndicated information feeds, Web-scraped news and social postings, and corporate documents, for person and company names, ticker symbols, flagged terms, and so on. Partner solutions from Kapow Software and ISYS Search may prove of great value in acquiring and processing these textual materials in preparation for UDA full-text indexing and search. Another Sybase partner, BMMSoft, offers a complete end-to-end data extraction to regulatory hold and reporting package called EDMT Server that is built on Sybase IQ and UDA. The BMMSoft EDMT package is designed for regulatory and compliance environments and for eDiscovery and litigation scenarios.

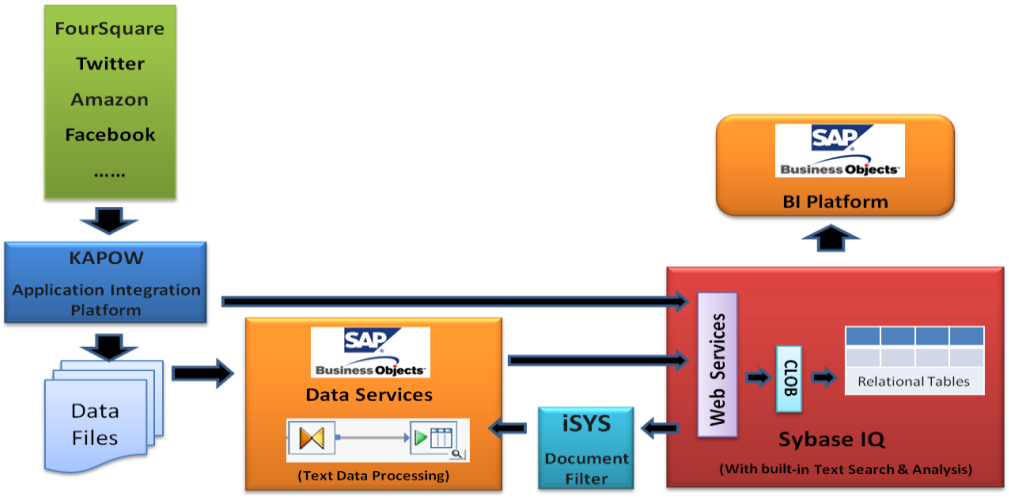
Technical support analyses

This final scenario is actually a report of the Sybase support organization's experience with a tracking database, built with Sybase ASE 15.x DBMSes, containing millions of rows of real-time and historical data dating back twenty years. The data consist of customer case communications, support-activity logs, and error logs with problem reproductions, etc.

Replication Server 15.5 was set up to populate Sybase IQ database with UDA set up for TEXT indexing. The team ran simple, multiterm, Boolean, prefix, and proximity SQL search queries. They experienced 20x to over 1000x execution-time reductions with IQ over ASE, showing the effectiveness of replicating real-time operational data to a Sybase IQ analytical DBMS for search and query.

A unified processing and analysis system

The figure below, provided by Sybase, shows key technologies from SAP, Sybase, and partners (Kapow Software and ISYS-search software) provides a generic workflow that is applicable to the use cases described above.



CC BY ND

Sybase IQ's Key Value Proposition

It is not difficult to explain Sybase IQ's unique text-analytics value proposition, the rationale for choosing Sybase IQ as a text-analytics platform.

Sybase IQ excels at handling Big Data, whatever its type – structured or unstructured – and whatever its source – the enterprise transactional and operational systems that fuel conventional BI, or data and unstructured information from online, social, and enterprise sources.

The Sybase IQ platform provides the capabilities, performance, scalability, and security essential for enterprise data analyses via its industry-leading columnar, parallel, analytical database architecture and partner solutions that extend Sybase IQ data acquisition, processing, and analysis possibilities.

In fact, Sybase IQ is the only leading analytical database platform to provide integrated text processing and analysis functions. These capabilities rely on Sybase IQ's long-standing text-management ability and take advantage of Sybase IQ's columnar architecture, which minimizes I/O and supports aggressive data compression, including of text columns. Sybase IQ's capabilities are extended by the Unstructured Data Analytics option and complemented by Kapow Software's Web- and enterprise information acquisition suite, by ISYS Document Filters for document processing and text extraction, and by SAP BusinessObjects Data Services, including Text Data Processing. The platform provides comprehensive support for essential text-data management, search, query, analysis, and visualization.

Sybase IQ in-database analytics capabilities that cover both text and conventional fielded data and enable unified analysis, providing the ability to derive insights from data that draw from the richness of text, the ability to discern the *Voice of the Customer* and other qualitative information that cannot be found in transaction records.

The sum of these capabilities constitute Sybase IQ's key value proposition, the analytical database platform of choice for enterprise text-data analytics.

CC BY ND

About

Seth Grimes

White paper author Seth Grimes is an information technology analyst and analytics strategy consultant. He is contributing editor at TechWeb's *InformationWeek* and founding chair of the *Text Analytics Summit* and the *Sentiment Analysis Symposium*. He served as an instructor for *The Data Warehousing Institute (TDWI)* and text analytics channel expert at the *Business Intelligence Network* and is the author of numerous articles and reports covering a range of topics in BI and analytics and their business applications.

Seth has worked with database, BI, and decision-support applications and users for over 25 years. He founded Washington DC-based Alta Plana Corporation in 1997. He consults, writes, and speaks on information-systems strategy, data management and analysis systems, industry trends, and emerging analytical technologies.

He is the author of the 2010 Sybase sponsored white paper, Text Analytics in the BI Ecosystem, available online at *http://bit.ly/SybaseTA*.

Seth can be reached at *grimes@altaplana.com*, +1 301-270-0795. Follow him on Twitter at *@sethgrimes*.

Sybase

For 25 years, Sybase has been a leader in developing and expanding innovative database technology. Since its founding in a Berkeley, Calif., home in 1984, Sybase has earned the trust of many of the world's leading companies for its ability to manage information and deliver unsurpassed levels of data reliability and security. Today, Sybase is the largest enterprise software and services company exclusively focused on managing and mobilizing information. With its global solutions, enterprises can extend their information securely and make it useful for people anywhere using any device.

Sybase

One Sybase Drive
Dublin, CA 94568
Phone: +1 925-236-5000

Visit www.sybase.com.

CC BY ND