TeraText DBS Frequently Asked Questions (FAQ)

TeraText DBS Frequently Asked Questions (FAQ)

What is the TeraText® Database System (DBS)?

The TeraText® Database System (DBS) is a non-relational text database and search tool that scales to handle very large collections of unstructured and semi-structured text. It provides capabilities for storing, indexing, searching, and retrieving documents and any associated metadata. The TeraText DBS has been optimized for XML and is often used as an XML database.

How long has the TeraText DBS been in existence?

The technology was first developed by researchers at the Royal Melbourne Institute of Technology (RMIT) in Melbourne, Australia, in the early 1990s. The first commercial sales occurred in 1994. SAIC has been the exclusive distributor of the TeraText DBS in North America since 2000.

What are the key discriminating features of the TeraText DBS?

The TeraText DBS differentiates itself in the marketplace in the following areas:

- Performance – high-performance throughput; can load and search text simultaneously

- Scalability – scales to handle terabytes of information via a Z39.50-based distributed architecture. Servers can be "hot-added" without bringing down the system

- Data Compression – minimizes storage requirements and speeds data transfers

- Data types – handles over 255 types including plain text, SGML, XML, office documents and binary

- Unified access – searches disparate text collections using a single logical view

- Multilingual – full Unicode support for storing, indexing and searching

- Security – role based access control at the database, document, and field levels

What are typical applications of the TeraText DBS?

Applications developed using the TeraText DBS include knowledge bases, legislation systems, Web portals, technical documentation systems, digital libraries, and email management systems. The product is particularly well suited for high-volume, high-throughput, workflow environments in which documents need to be quickly loaded and indexed for searching. TeraText typically outperforms relational databases in these environments. The TeraText DBS excels at finding information in massive amounts of relatively unstructured data.

How scalable is the TeraText DBS?

The TeraText DBS is designed to accommodate relatively small text collections of less than a gigabyte to those in the multi-terabyte range. This scalability is made possible by a distributed or federated architecture that is based on Z39.50, an international standard for searching and retrieving text across a wide area network. As a document collection grows over time, additional servers can be added, each containing a portion of the collection as a separate TeraText database. As an example, one of our customers started with a database of 50 GB and has steadily scaled it to approximately 4 TB.

What kind of performance does the TeraText DBS deliver?

Performance is highly dependent on the configuration of the TeraText database and the computer platform it runs on. As an example, the TeraText DBS on a single 4-CPU Pentium® IV device running Microsoft® Windows® XP can load and full text index about 1 MB of text per second. It can support searching of over 47 million document pages per second with 350+ named query users per CPU. Using TeraText's compression technology, 1 TB of original XML data is typically stored (with indexes) on less than 1 TB of disk.

How does the TeraText DBS differ from other XML database software?

The TeraText DBS contains unique features such as the ability to search across disparate text collections stored on local or remote servers. Its distributed architecture provides exceptional scalability and performance. It enables loading of entire XML documents as well as the creation of "virtual fields" which are dynamically generated at load time from subsets or combinations of specific XML elements. It is able to load and index new documents in real time while the database is live and supporting queries. It provides a rich suite of query features such as fuzzy matching and database scanning. It accommodates not only XML but also plain text and numerous binary document formats. And the TeraText DBS security model is best-of-breed, providing access control at the database, document and XML field levels.

How does the TeraText DBS differ from relational database products?

Relational databases are very good for storing highly structured information that fits nicely into tables, rows, and columns. Semi-structured and unstructured text, on the other hand, is more difficult to store in a relational database because it means you either have many different tables (many joins and slow retrieval times) or a single table with many null columns (storage inefficiencies). XML documents present particular challenges because they have a tree-like structure whereas the relational model is grid-based. Mapping XML DTD's or schemas to a relational model is often non-trivial. Because the TeraText DBS was designed to accommodate XML and other types of textual data, it provides exceptional loading and querying performance of these data compared to relational databases. It also preserves the native format of XML documents, allowing one to perform complex queries that exploit the hierarchical structure of the documents.

What is unstructured, semi-structured, and structured text?

Structured text typically refers to data with well-defined attributes that lends itself well to storage as rows and columns in a relational database. Unstructured text comprises free-form text of arbitrary sizes and types. Examples include letters, office memos, reports, and Web pages. Semi-structured text lies between these two extremes. It has some structure, but is not rigidly structured. Examples of semi-structured text include catalogs, email messages, technical documentation, and XML documents. It is estimated that approximately 80 percent of an organization's knowledge is stored in unstructured and semi-structured documents.

Does the TeraText DBS manage images and other non-textual data?

Yes. The TeraText DBS has a binary data type that can store be used to store images, spreadsheets, office documents, PDF's, CAD files, and many other types of documents (up to 2 gigabytes in size). A document conversion plug-in is available that can convert up to 255 different types of documents to XML or plain text.

What search capabilities are provided?

Search capabilities include fuzzy matching, proximity queries at the word, sentence, and paragraph levels, custom term weighting, boolean operators, ranking, sorting, field-based queries, wild cards (truncation) and saving of result sets for "search within" functionality. Queries can be saved and rerun at specified intervals for automated retrieval of new information. The TeraText DBS also provides a "scanning" feature that allows one to browse through indexed fields (such as subject lines in email messages) as an alternative approach to querying for discovering information.

What are the security features?

The TeraText DBS provides role-based access to data at the field, record, and database levels. This enables an administrator to restrict access to sensitive data down to the level of specific XML nodes. Other security features include support for Lightweight Directory Access Protocol (LDAP) to manage user and group information, the Generic Security Service (GSS) to protect client connections via standard security technologies, and Secure Sockets Layer (SSL) to handle encrypted sensitive information.

Is an application programming interface (API) provided?

Yes. The TeraText DBS provides a feature-rich application development environment that includes an extensive suite of libraries in Java™, C# (.NET), C++, and ACE, TeraText's own built-in scripting language. Programming modules are available for tasks such as searching, scanning, presentation and sorting of results, and creating "task packages" to access stored queries. Other modules are provided for database initialization, updating and deleting records, access control, LDAP integration, database backup, rebuilding indexes and error handling.

What computer platforms does the TeraText DBS run on?

The software runs on Microsoft® Windows® 2000 and XP, SUNSM Solaris™ 9.0 or higher, and several flavors of Red Hat® Linux®.

TeraText® DATABASE SYSTEM