Keeping Up With Terabytes of Continuously Updating Data

How do you keep up with terabytes of constantly updating data? The answer is surprisingly easy. Simply deploy a search engine with indexed search and set it for automatic index updates. Here's how that works.

A search engine like dtSearch® instantly searches terabytes through indexed search. While you can do an unindexed search, or search without first building an index, that tends to be much slower. In contrast, indexed search across terabytes can get you search results in an instant.

To index, just point to the directories and the like you want to cover, and the search engine will go off on its own and do the rest. If you are using dtSearch, no need to tell it what types of files, emails, etc. it is working with; the product line figures that out on its own. And once you have one or more up to terabyte-size indexes, the products can instantly perform over 25 different kinds of full-text and metadata searches, displaying retrieved items with highlighted hits.

The indexer can automatically detect if files or other items have been edited, added or deleted since the previous index job. And you can set the application to do automatic index updates as often as you like using the Windows Task Scheduler. That way, whenever something changes, the search engine will remain on top of the situation, letting you instantly retrieve the most recent items.

In fact, the product line can update an index without blocking out searching. Even concurrent multithreaded searching can proceed while an index refreshes. There is no reason, therefore, not to keep indexes updated, keeping you up-to-date with your evolving data. Indexed search can further cover a very wide range of data types.

The products automatically work with popular "Office" file formats, like PDF and Microsoft Word, Excel, Access, PowerPoint and OneNote. The product line can work with these in a standalone format or as part of a compressed archive like ZIP or RAR. The products also work with popular email formats including Outlook and Exchange. And they work with web-based formats such as HTML, XML and other online data. The developer products further work with databases like SharePoint, SQL and NoSQL, including database metadata and referenced or BLOB files.

Documents saved with a mismatched extension, like a PDF saved with a .DOCX extension, are not an issue. The products determine the relevant file type by looking inside each file, rather than relying on the document extension. And not only does the search engine automatically recognize and support free-standing files, but also multilevel nested attachments. If you have an email with a ZIP file attachment and embedded in that ZIP file is an Access database, and embedded inside the Access database is an Excel file, the product line will parse the whole thing—even if the embedded Access and Excel files are mislabeled with say .DOCX or PDF extensions.
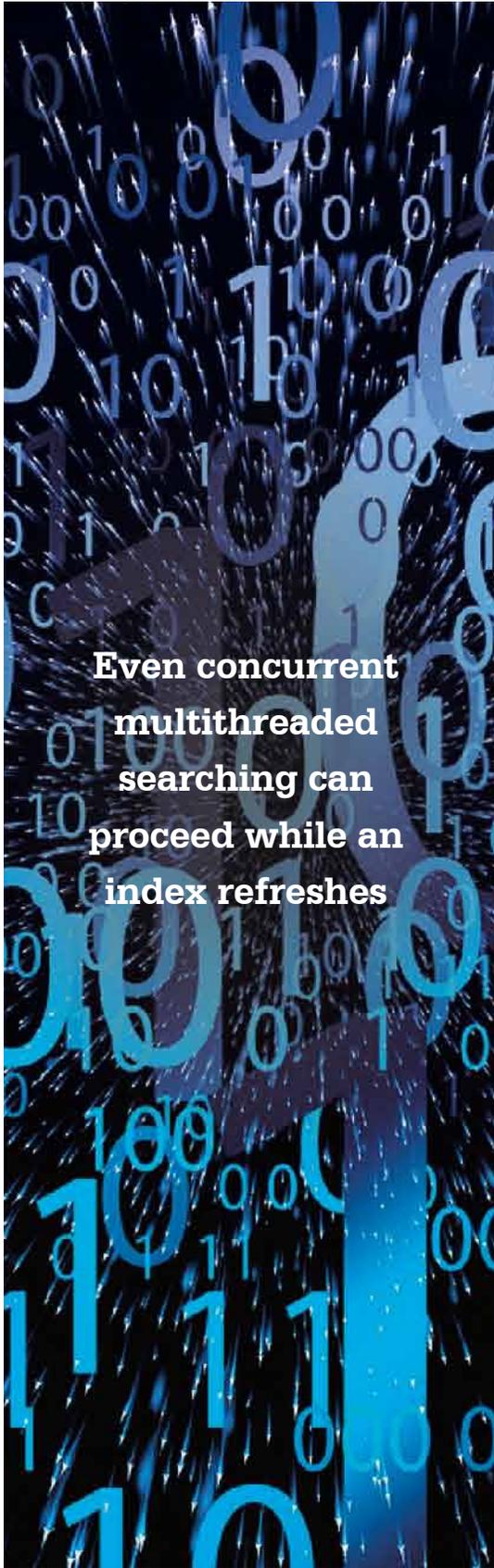
Even concurrent multithreaded searching can proceed while an index refreshes

For working with continuously updating data, there is another option as well. Normally, when you build an index, the products return to the original files, databases, emails, etc. to display retrieved items with highlighted hits. But there is also a caching option. With caching on, an index stores the full content of the original files, databases, emails and the like inside the index along with the core search index information. That way, if an item is slow to retrieve or goes completely offline, the search engine can still immediately display it with highlighted hits.

Regarding emails, the products offer two general options for indexing Outlook and Exchange data. If the emails are "live" in Outlook, the products can access the emails via Microsoft MAPI. Alternatively, the products can access the emails directly just like ordinary files, bypassing MAPI altogether. The direct-access option is more efficient if you have a choice.

Regardless of how you access the emails, the search engine can index not only the full text of the email and all metadata, but any nested attachments. And the products can also copy a select item out of an email archive or a ZIP or RAR archive—or both types of archives together—just by clicking on the item and telling the application to copy it.

Working with international language text is also not an issue. Nearly all of the 25+ search options work with text in any of the hundreds of international languages supported by Unicode. This encompasses not only European languages, but also right-to-left languages like Hebrew and Arabic along with double-byte character text like Chinese, Japanese and Korean.

Fuzzy searching, for example, which looks for whatever word you type with typographical deviations adjustable from 0 to 10, can operate not only with English but also other international languages. Fuzzy searching is great for sifting through items like emails which can have frequent misspellings or OCR'ed PDFs which might have OCR errors. The products can even locate items like credit cards in international language text or generate and search for hash values in international language files.

dtSearch has enterprise and developer products that run "on premises" or on cloud platforms to instantly search terabytes of "Office" files, PDFs, emails along with nested attachments, databases and online data. Because the product line instantly search terabytes with over 25 precision search options, many customers are Fortune 100 companies and government agencies.

But **anyone** is welcome to go to dtSearch.com to download a fully-functional 30-day evaluation version to instantly search terabytes in a standalone capacity, or in a concurrent-search capacity in a shared network or online environment.

You can set the application to do automatic index updates as often as you like using the Windows Task Scheduler