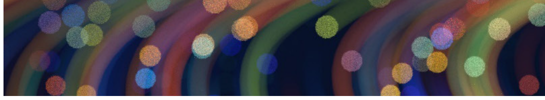


Multithreaded Enterprise Search

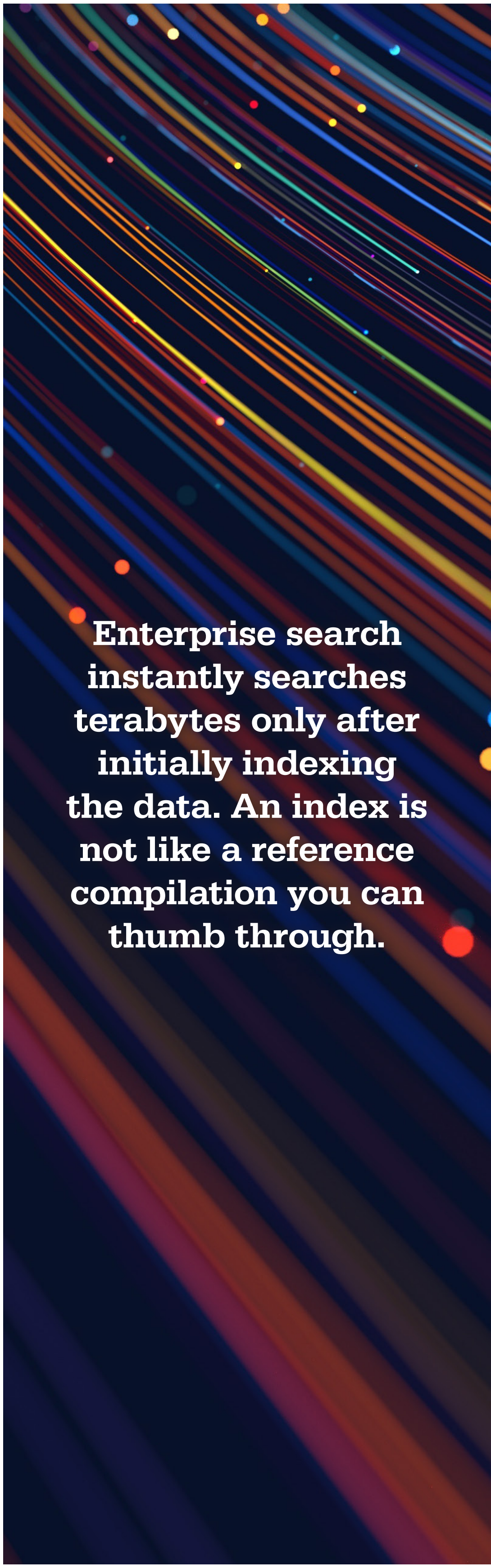
Article contributed
by dtSearch®

Today's focus is multithreaded enterprise search. With dtSearch®, for example, multithreaded processing makes its first appearance during the indexing phase. Enterprise search instantly searches terabytes only after initially indexing the data. An index is not like a reference compilation you can thumb through. Rather, an index is an internal mechanism for the sole purpose of letting end-users instantly and concurrently query large volumes of data. A single dtSearch index can hold up to a terabyte of text, and there are no limits on the number of indexes a search request can encompass.

Indexing couldn't be easier. Just tell the indexer the folders and the like to cover, and the indexer will take it from there. Indexed data can include email archives, document folders and web-based data formats. The indexer automatically works with any combination of local and remote files so long as the indexer can see remote documents like SharePoint attachments as well as OneDrive and DropBox files through the Windows folder system.

To correctly parse a file, the indexer needs to figure out its exact file format. By itself, the parsing mechanism or "document filters" can determine the exact file type using each file's binary format. This makes file identification a lot more foolproof than relying on the file extension. A PDF with a .DOCX extension, a OneNote file with a PowerPoint extension, an Access database with an Excel extension, etc., will not affect parsing.

You can further have an email with a ZIP or RAR attachment containing an Excel spreadsheet with a Word document recursively embedded inside and the indexer will support the whole thing, including all text and metadata. And dtSearch works with Unicode to cover hundreds of international languages. A file can go from English to another European language to double-byte Chinese, Japanese and Korean text to right-to-left text, and Unicode and dtSearch will track the whole progression.



Enterprise search instantly searches terabytes only after initially indexing the data. An index is not like a reference compilation you can thumb through.

Reprinted with permission of USA Daily Standard, usdailystandard.com

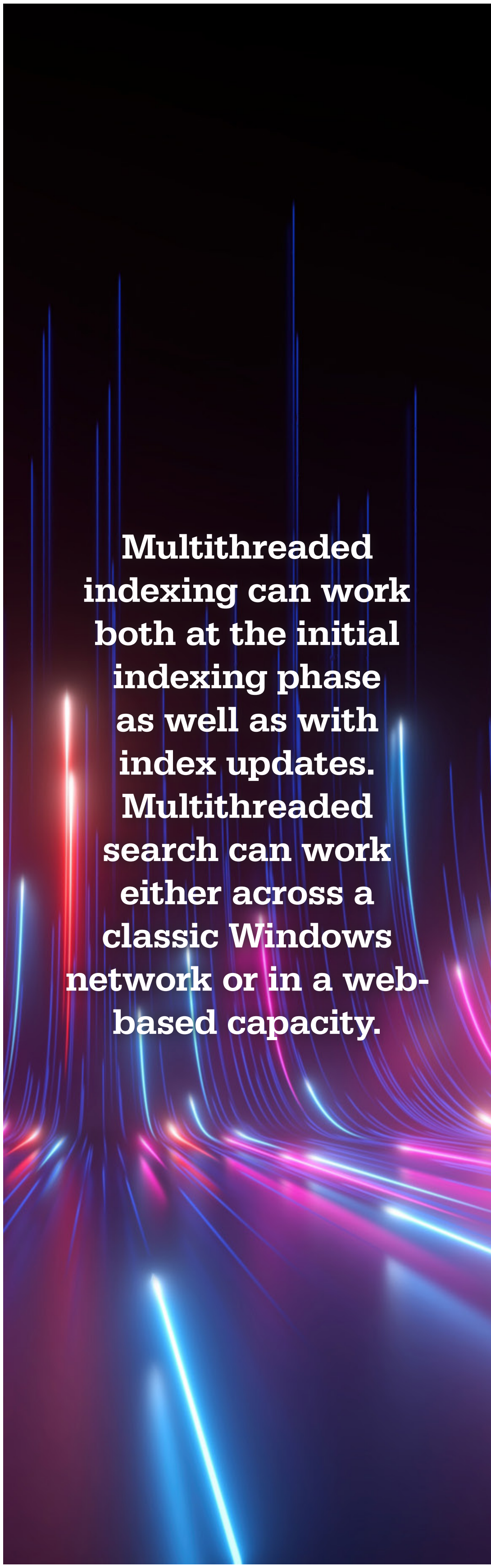
dtSearch has a multithreaded 64-bit indexer which can greatly improve indexing performance, with speed increases about 6 times faster on newer hardware. Multithreaded indexing can work both at the initial indexing phase as well as with index updates. Multithreaded search can work either across a classic Windows network or in a web-based capacity. Web-based searching can operate either through a local server or through a remote web server such as on Azure or AWS. In all cases, each search thread operates independently.

dtSearch has more than 25 search types all of which can work in a multithreaded environment. Natural language “all words,” “any words” or “exact phrase” are basic search requests. Take *Largish Company merger with Hugely Corporation*. An “all words” search would retrieve only files containing all of these words. An “any words” search would find any item that contains even just one of these search terms. An “exact phrase” search would look solely for files with this precise sequence of words together.

These include Boolean (and/or/not) and proximity searching. These options let you enter a more precision search request like *Largish Company and Hugely Corporation* within 12 words of *merger or legal department* in a file with no mention of *antitrust* in subject metadata. You can also add on a number or a numeric range element or a date or a date range component to a search request. The latter will even find hits across different date presentations, like *Dec 1, 2024* and *12/1/24* in a date or date range searching including *December 1, 2024*.

Stemming can find different variants on the same root word, like *merge*, *mergers* and *merging* in a search for *merger*. Fuzzy searching adjusts from 1 to 10 to sift through typographical or OCR errors like *department* mistyped as *departnent* in an email or mis-OCR’ed as *departmamt*. Concept or synonym searching can find *business* for *company*. dtSearch can even flag valid credit card numbers across data or generate hash values for each file and selectively search on them.

Article contributed
by dtSearch®



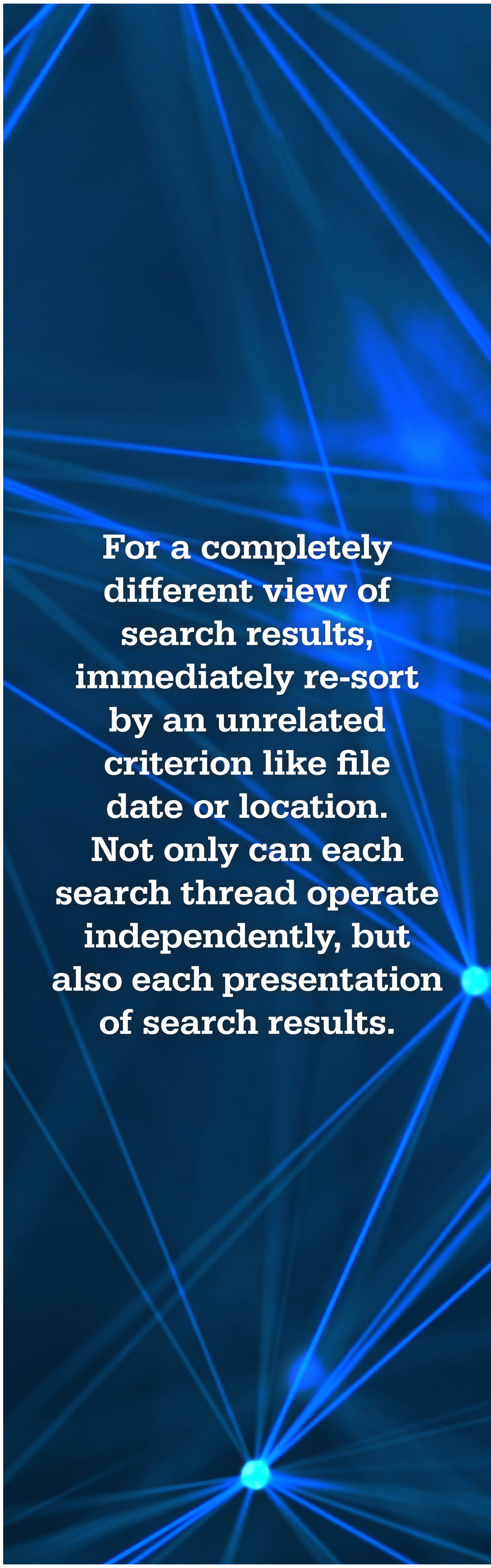
Multithreaded indexing can work both at the initial indexing phase as well as with index updates. Multithreaded search can work either across a classic Windows network or in a web-based capacity.

Reprinted with permission of USA Daily Standard, **usdailystandard.com**

By default, dtSearch will apply vector-space relevancy-ranking across all indexed data. Take an "any words" search for *Largish Company merger with Hugely Corporation*. If *company* and *corporation* are all over the data but *Largish* and *Hugely* much rarer, *Largish* and *Hugely* will get higher relevancy ranks and items with the densest mentions of these coming out on top. Or add in your own variable term weighting, giving *Largish* a positive weight of 3 and *Hugely* a negative weight of 6 but a positive weight of 9 for an occurrence in certain metadata or near the top or bottom of a file.

For a completely different view of search results, immediately re-sort by an unrelated criterion like file date or location. Not only can each search thread operate independently, but also each presentation of search results. After a search, end-users can view the full text of retrieved items with highlighted hits for convenient browsing.

Article contributed
by dtSearch®



For a completely different view of search results, immediately re-sort by an unrelated criterion like file date or location. Not only can each search thread operate independently, but also each presentation of search results.

Reprinted with permission of USA Daily Standard, **usdailystandard.com**