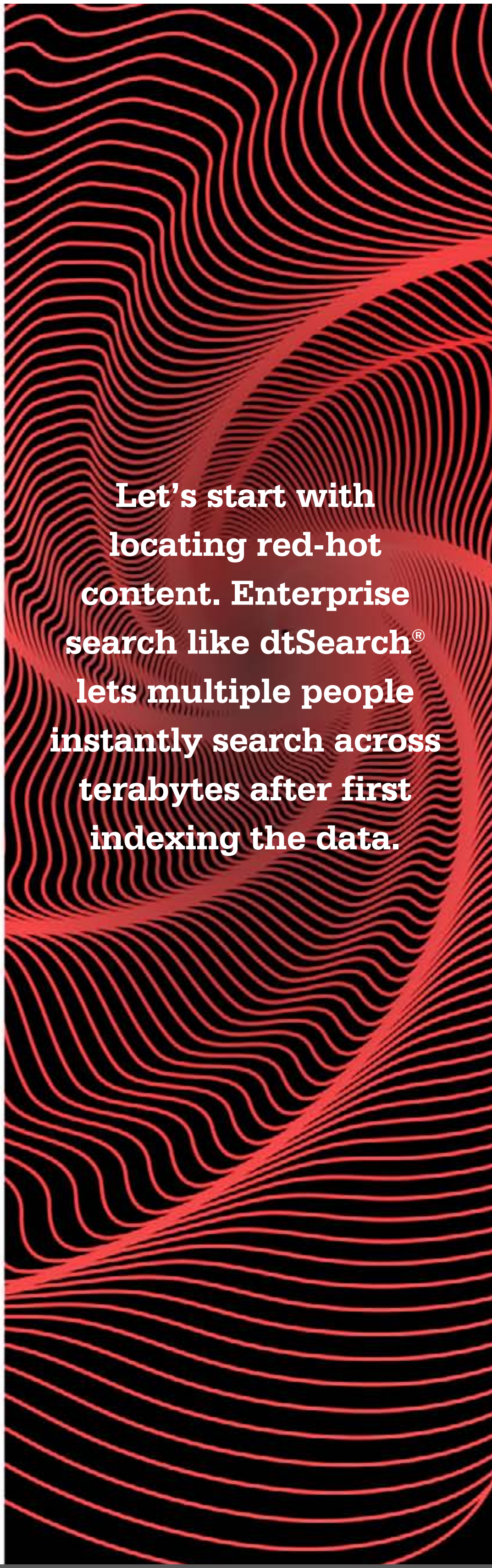
Pinning Down Red-Hot Content

Article contributed by dtSearch®

Since we're still in the thick of summer, I wanted to address practically too-hot-to-handle red-hot data. Enterprise search can locate the red-hot content you are looking for amidst terabytes of other organizational data. Looking ahead, building an application around the software developer kit or SDK can limit the retrieval of red-hot data to authorized internal users only, while making such data inaccessible to the wrong internal users.

Let's start with locating red-hot content. Enterprise search like dtSearch® lets multiple people instantly search across terabytes after first indexing the data. Indexing is easy. Simply point to the folders and the like to index, and the dtSearch indexer will take it from there. In doing so, the indexer will go through each item and figure out if it is a Microsoft Word document, an Excel spreadsheet, an Access database, a PowerPoint, a OneNote file, a PDF, an email file, a compressed archive, etc. and parse that item accordingly.



Let's start with locating red-hot content. Enterprise search like dtSearch® lets multiple people instantly search across terabytes after first indexing the data.

The indexer uses the information inside each binary file itself to figure out the format. That way, a Word file ending in a .PDF file extension or a PDF ending in a .ONE file extension will not trip up the indexer. The indexer will also support all text and metadata for each item, even finding hits in obscure metadata that you may not even have realized was there if you looked at the item in its native application.

The indexer will also cover text that may be camouflaged in a native application view, like hot pink text against a hot pink background or cool blue text against a cool blue background. Additionally, the indexer can make its way through recursively nested formats. If there is an email with a ZIP or RAR attachment, and inside is a Word document that itself contains an Excel spreadsheet, the indexer will span everything. Further, so long as remote files like Office 365 files or SharePoint files present as part of the Windows folder system, the indexer can index and search them just like local content.

dtSearch also lets you use the Windows Task Scheduler to manage index updates automatically to accommodate added, deleted or modified content. Updating an index does not affect individual or concurrent searching so queries can continue uninterrupted while the index updates. Indexing is resource intensive, but concurrent searching, whether across a standard Window network or from an Internet or Intranet configuration, is very resource light. In fact, web-based searching from the cloud like AWS or Azure or from an on-premises Internet or Intranet server can proceed statelessly, with no built-in limit on the number of simultaneous search threads.

Reprinted with permission of Daily Business Journal, dailybusinessjournal.com

dtSearch has over 25 different search options. Search requests can range from simple unstructured natural language to intricate Boolean and proximity word and phrase full-text and/or metadata-specific formulations. Stemming covers different endings on the same root word. Concept searching can expand a search request to cover synonyms of search terms. Fuzzy searching adjustable from 0 to 10 can sift through typographical or OCR deviations, letting a search for *Red-Hot Radioactive* find not only *Red-Hot Radioactive* but also *Red-Hot Radioactine* that someone may have mistyped in an email.

dtSearch can also search for numbers and numeric ranges as well as dates and date ranges in specific metadata or in the full text. Date and date range searching can even automatically pick up different date formulations like *Jan 13, 2023* versus *1/13/23*. Speaking of red-hot data, dtSearch can even find any credit card numbers that may be lurking in a dataset. Unicode support covers hundreds of international languages, including right-to-left languages like Hebrew and Arabic and double-byte Asian text like Chinese, Japanese and Korean. A single file or email can cycle through multiple international languages, and Unicode and dtSearch will cover the whole progression.

dtSearch's default relevancy-ranking uses a so-called vector-space algorithm that gives more common indexed search terms a lower relevancy ranking and less common indexed search terms a higher relevancy ranking. Files with the densest and rarest search terms get the highest relevancy ranking. You can also override the default rankings through custom positive or negative variable term weighting across all text, or with enhanced weight at the top or bottom of files or in certain metadata. Or instantly re-sort by some completely different metric like filename or file location. Whatever the sorting, you'll get a full copy of retrieved items with highlighted hits for convenient navigation.

If an organization requires differential security settings across the data, the system administrator can make available different indexes for different end-users. Going one step further, an application built around the dtSearch Engine SDK can granularly filter content using any combination of SQL, NoSQL or SharePoint metadata; file metadata; metadata added "on the fly" while indexing; or full-text references. That way, if *ProjectXYZ* suddenly becomes red hot, an application embedding the dtSearch Engine can mandate that any mentions of *ProjectXYZ* in the full text or metadata remain "eyes only" to the correct people.

About dtSearch.® 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 dtSearch can instantly search terabytes with over 25 different search features, many dtSearch customers are Fortune 100 companies and government agencies. But anyone with lots of data to search can download a fully-functional 30-day evaluation copy from dtSearch.com

Article contributed
by dtSearch®

Speaking of red-hot data, dtSearch can even find any credit card numbers that may be lurking in a dataset.

Reprinted with permission of Daily Business Journal, **dailybusinessjournal.com**