Searching Terabytes – Just the Basics

Article contributed by dtSearch®

It is the beginning of a new school year. But while academics can be hard, below are *just the basics* that you need to instantly search terabytes. To start with, enterprise search can search terabytes using unindexed search or indexed search.

Unindexed search would be similar to rummaging through your sock drawer looking for your lucky "first game of the season" socks. If you rifle through the back left of the sock drawer, then the front left, then the back right and the front right, you know they'll eventually turn up. But now imagine that you need to rummage through not just one sock drawer but millions of sock drawers.

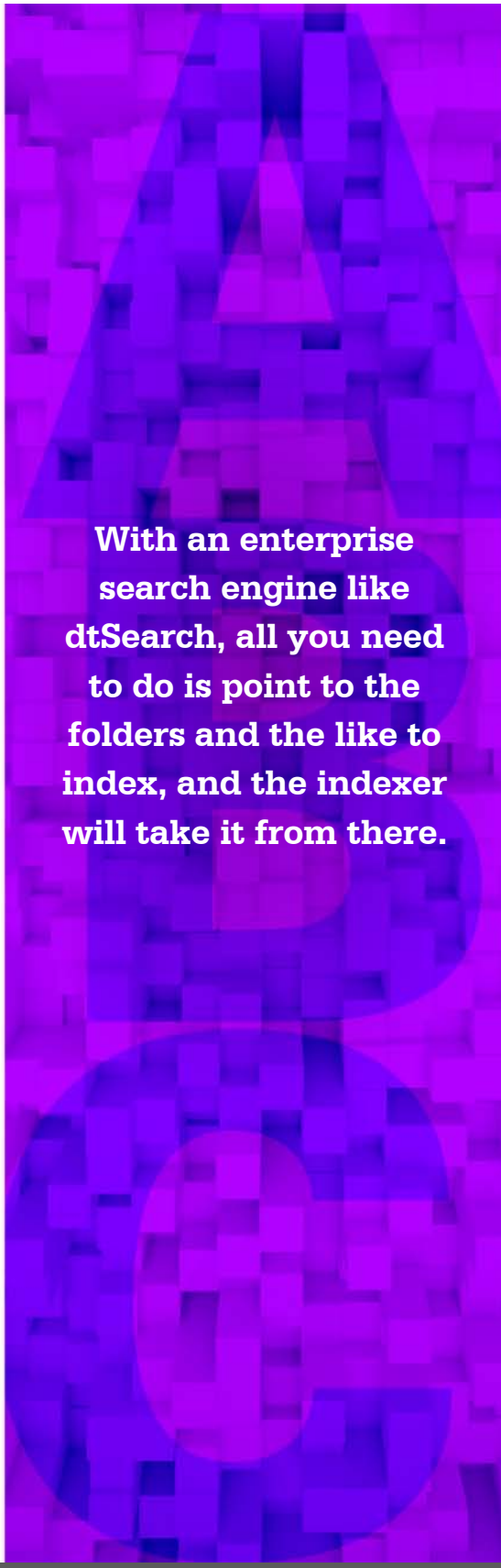
The alternative is indexed search. For your sock drawer, the index would include a quick description of each pair of socks and a mapping of where each fits in the drawer. For data, the index covers every unique word and number and the location of each. Just like an index of your sock drawer would speed up your ability to find your lucky socks, indexing enables enterprise search to span a lot more content faster, covering terabytes in an instant.

While your sock drawer isn't going to index itself, enterprise search can go through the underlying data on its own. With an enterprise search engine like dtSearch®, all you need to do is point to the folders and the like to index, and the indexer will take it from there. As a first step, the indexer will go through each item in the folders and figure out what type of file each is.

Parsing specifications for PDFs, Microsoft Word documents, Excel spreadsheets, Access databases, PowerPoints, OneNote files, email formats, etc. are all radically different, so the indexer has to get its file format determination right. It is quite easy, however, to save a Word document with a .PDF extension or an Excel spreadsheet with an Access database extension. For accuracy, the indexer needs to look inside each binary format to make its file type determination.

In going to the binary formats, the indexer can support not only individual files, but also multilayer nested formats. An email can include a ZIP or RAR attachment containing a PowerPoint with an Excel spreadsheet embedded inside and the indexer can cover everything. And while file display in an associated application view—like a Word document display inside Microsoft Word—can obscure certain text and metadata, the binary format view of a file renders everything clearly for the indexer.

For example, black text against a black background or white text against a white background that may all but disappear inside an associated application display is just plain text inside the binary format. And metadata that may take a huge amount of clicking around to find in an associated application view is likewise immediately available inside the binary format. The indexer can even



With an enterprise search engine like dtSearch, all you need to do is point to the folders and the like to index, and the indexer will take it from there.

handle remote data. As long as Office365 files or SharePoint attachments, for example, present as part of the Windows folder system, the indexer can work with them just like any other files.

Indexed search can instantly span terabytes with one or more search threads. Indexing is resource-intensive. But concurrent searching, whether across a classic Windows network or from an "on premises" web server or a cloud server like Azure or AWS, is very resource-light. Internet or Intranet searching can operate statelessly, with no built-in limit on the number of concurrent search threads. While it may get crowded if multiple people simultaneously rummage through the same sock drawer, enterprise search threads can proceed independently and instantly.

What happens when data changes? You can set the enterprise search indexer to automatically update indexes using the Windows Task Scheduler to accommodate file additions, deletions or modifications as frequently as you want. Updating an index does not stop concurrent searching so there is no reason not to keep indexes current.

Enterprise search supports over 25 different search options, ranging from "all words" or "any words" natural language searching to highly complex Boolean and proximity word and phrase full-text and metadata search formulations. Concept searching expands to similar concepts. Fuzzy searching adjustable from 1 to 10 sifts through mistypings or mis-OCRs like *sock draver* instead of *sock drawer*.

Date recognition finds dates or date ranges in full-text or metadata, including automatically extending to different date formats. Credit card searching can identify any credit card numbers that may appear in indexed data. Unicode support covers hundreds of international languages, including right-to-left languages and double-byte character text.

For sorting files, vector-space relevancy-ranking is the default. In a search for *sock or drawer*, if drawer appears in thousands of files but *sock* only appears in a handful, then files mentioning *sock* would get a higher relevancy ranking, and files with the densest *sock* mentions would get the highest ranking. Custom positive or negative variable term weighting is also an option either across all data, or with a higher weight at the top or bottom of files or in certain metadata. Or instantly re-sort by an unrelated criterion like file date or file location. Enterprise search will display a full copy of retrieved files with highlighted hits for convenient navigation.

In sum, time to make instant concurrent searching across terabytes of data a reality for your organization. A download link follows below.

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®