Pot of Gold at the End of the Rainbow Meets Big Data

Article contributed
by dtSearch®

Find out how to leverage indexed search and make the most of Big Data.

If you are looking for a pot of gold at the end of the rainbow, a treasure map is your magic ticket. If you are digging for gold in Big Data, indexed enterprise search is your express token. You may have to cajole more than a few leprechauns to get a treasure map to the end of the rainbow. But indexed search across terabytes is instant and easy, with sufficient gold nuggets for everyone.

Big Data's "treasure map". Enterprise search engines are a different animal from the usual online search engines. Enterprise search offers two ways to comb through data: indexed search and unindexed search. Unindexed search iterates through every word of each file, email and the like, making the process slow, particularly across millions of items.

In contrast, indexed search queries an existing index, not the underlying data. In indexing, the search engine goes directly to the binary format of everything, bypassing the need to retrieve files in their associated applications (email programs, PDF viewers, web browsers for online formats, Microsoft Word, Access, Excel, PowerPoint, OneNote, etc.). Because indexing extracts and organizes all the critical information in advance, indexed search can span terabytes in an instant.

So how do you get an enterprise search engine to generate a search index? All you have to do is point to the folders, email repositories and the like to cover, and the search engine will take it from there. Once the index is done, the gold nuggets appear.

Gold nugget #1: The structure of the index enables multiple concurrent users to crisscross the index with search threads operating independently. Online, indexed search can run in a completely stateless manner, making it very easy to scale. And updating an index to add new data does not affect searching, so instant concurrent searching across a network or online can continue even while indexes automatically update themselves.

Gold nugget #2: The search index has sufficient information to support over 25 different search features. These range from

If you are digging for gold in Big Data, indexed enterprise search is your express token.

Reprinted with permission of Spiceworks, spiceworks.com

free-form natural language search requests to highly structured phrase, Boolean (and/or/not), and proximity search requests: *("pot of gold" w/8 rainbow) and (leprechaun or "good luck") and not ("lottery win" or jackpot)*. Concept searching finds synonyms and related words, like arc for rainbow. Metadata-specific searching lets you limit a search or a portion of a search to specific metadata only.

Gold nugget #3: Fuzzy searching sifts through misspellings, finding *lepreckauns* for *leprechauns*, resulting from either a typo in an email or an OCR mistake in a PDF.

Gold nugget #4: Unicode support enables searching across hundreds of different international languages. These include European languages, right-to-left languages like Arabic and Hebrew, double-byte Chinese, Japanese and Korean text, and many more. Indexed search features like fuzzy searching can work with English text or any of these international languages. And Unicode support also covers emojis 😊

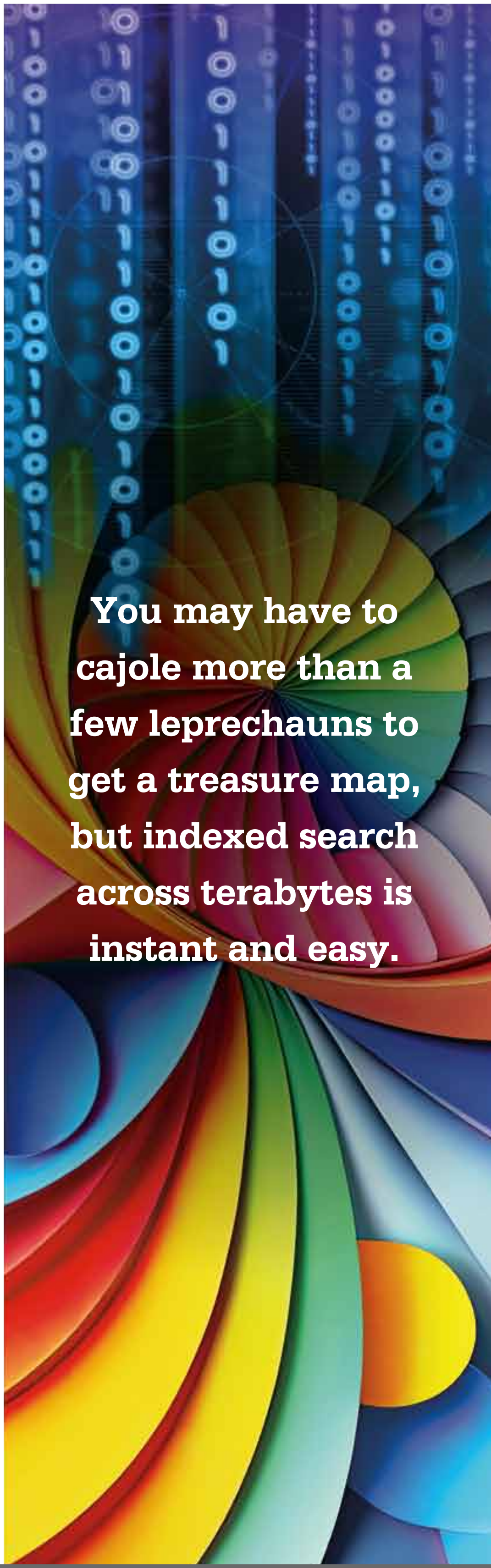
Gold nugget #5: In addition to word-based retrieval, indexed search encompasses numeric-oriented queries for numbers or number ranges as well as dates and date ranges, even automatically across different formats like February 15, 2023, and 2/15/23. Search also covers numeric expressions and can even identify valid credit card numbers that may appear in indexed data. Indexing can also generate hash values for all files and then enable searching on those values.

Gold nugget #6: Regardless of the type of searching, indexed search makes it easy to go through search results with numerous relevancy-ranking and other sorting options. The default vector-space ranking assigns a relevancy score to items based on the rarity and density of search terms in the underlying indexed data. If *luck* is prevalent in indexed data, but *leprechaun* appears in only a handful of items, *leprechaun* mentions would get a more elevated relevancy score. And items with the densest *leprechaun* mentions would get the highest ranking.

In addition to the default vector-space ranking, enterprise search also provides options for customizing ranking by assigning positive or negative weighting to terms regardless of their frequency. This variable term weighting can apply across all indexed data or positionally to terms appearing at the top or bottom of items. Variable term weighting can also apply only if terms appear in specific metadata.

Beyond relevancy, the search engine lets you choose a completely different method of sorting as a different window into search results. For example, you can just click to instantly re-sort by file date or file location. However, the

Article contributed
by dtSearch®



You may have to cajole more than a few leprechauns to get a treasure map, but indexed search across terabytes is instant and easy.

Reprinted with permission of Spiceworks, spiceworks.com

search engine ranks search results, and it can display a full copy of retrieved items with highlighted hits for convenient review.

The above are some obvious nuggets of gold resulting from indexed search. But indexed search includes some more obscure treasures as well. These less apparent nuggets stem from the fact that the search engine builds its index by accessing binary formats while bypassing files' associated applications.

Gold nugget #7: You can't fool a search engine with a mismatched file extension. Say you have an email saved with a .ONE extension or a PDF saved with a .DOCX extension. While that might trick someone looking at files in the file system or viewing files in their associated applications, it has no effect on the search engine. This has no impact because the indexer figures out the relevant file type from the binary format itself. The file extension is not relevant to this process.

Gold nugget #8: The search engine can see right through multi-level nested files, like an email with a ZIP or RAR attachment containing a Word file that itself embeds an Excel spreadsheet. Again, the fact that the search engine accesses files in their binary format makes this possible.

Gold nugget #9: There is no "hidden" metadata. Metadata that may take a ton of clicking around to find in a file's native application is immediately apparent in the binary format.

Gold nugget #10: There is no "hidden" text. Sometimes a file creator or editor will try to hide text by having it blend in with the background color inside the file's associated application. But gold writing against a gold background or white writing against a white background is just straight text from the perspective of the binary format. One side note: if a file has tracked changes that have not been fully "accepted," these, too, are searchable, even if they may not appear by default in the file's associated application.

Gold nugget #11: The search engine can flag pesky "image only" PDFs during the indexing process. You know when you have a PDF file that looks like an ordinary PDF from the outside, but when you try to copy and paste text from it, nothing emerges? That is likely an image-only PDF. The search engine can flag that so that you know you need to run that file through an OCR program like Adobe Acrobat to copy and paste from it – and to search it.

Whether or not you find a leprechaun with a treasure map to the end of the rainbow, get maximum gold from your Big Data with indexed enterprise search.

Article contributed
by dtSearch®

Reprinted with permission of Spiceworks, spiceworks.com