

Bulletin

The FAST Track to Industrial Strength Search

Analysts: Susan Feldman and Chris Sherman

IDC Opinion

What are the crucial issues for the enterprise to consider when upgrading search systems to handle the relentlessly increasing demands of employees and customers?

Once-adequate intranet or basic site search tools are increasingly overwhelmed, not only by the massive quantities of information they must sift through but also by new formats and new user demands that far exceed capabilities considered adequate just a scant few years ago.

Raw horsepower and scalability, though important, are no longer adequate for most enterprises. What's needed is "industrial strength" search, with advanced linguistics, tunable relevancy algorithms, and flexible, scalable architecture. These powerhouse systems are going to be increasingly vital to the health and success of companies operating in today's brutally competitive information economy.

The quaint 20th century notion that a one-size search application fits all organizations has been undone by the relentless growth of mission-critical information within the enterprise. As the information economy continues to grow, organizations striving to keep their competitive edge will increasingly find that one size or flavor of search application does not fit all needs. IDC sees niches emerging for specific kinds of search in different circumstances:

- **Industrial strength search.** These powerhouse systems must scale to handle millions of documents and millions of queries. To handle these demands, subsecond response times are required, not only to satisfy the information needs of impatient users but also to discover and index new information the moment it is added, anywhere on any system within the enterprise. To further enhance usability, industrial strength systems should also offer proactive capabilities, filtering, alerting, and pushing new information that is custom tailored to the needs of each individual user. This kind of load should never be placed on the company intranet.
- **Question-answering systems.** Unlike traditional search engines, which provide a list of “results” that must be painstakingly examined to satisfy an information need, question-answering systems offer direct, unambiguous information — literally answering a user’s question. These types of systems are increasingly crucial for call centers, both to provide 24 x 7 self-help customer service and to better enable customer representatives to quickly dispatch telephone queries. In both cases, well-tuned question-answering systems can offer a profound, virtually immediate ROI, slashing support costs and freeing up knowledge workers for more productive activity. Question-answering systems can also lower costs and improve productivity within internal service functions such as human relations and benefits management.
- **Customized intranet search.** Rather than simply using out-of-the-box search applications included with the variety of server, database, and other applications installed on the intranet, search should increasingly be geared to the internal demands of the organization. Search should be capable of handling the diverse range of systems, including data stored in structured and unstructured formats. For some enterprises, this may mean that intranet search may exclude some extranet applications because of security concerns and because of the heavy demands that outward-facing Web sites may place on an intranet.

Quoting IDC Information and Data: *Internal Documents and Presentations*—Quoting individual sentences and paragraphs for use in your company’s internal communications does not require permission from IDC. The use of large portions or the reproduction of any IDC document in its entirety does require prior written approval and may involve some financial consideration. *External Publication*—Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2002 IDC. **Reproduction is forbidden unless authorized.**
For additional copies please contact Cheryl Toffel, 508-935-4389.

Check us out on the World Wide Web!

<http://www.idc.com>
Printed on recycled materials. ♻️

- **Email management.** Email, long regarded as a secondary, disposable information channel within the enterprise, has now assumed a vitally prominent role. Savvy organizations are awakening to the value of information squirreled away in many email messages, recognizing the payloads as an important aspect of corporate memory. External forces are also making organizations treat email with greater care. In response to corporate chicanery, new regulations require many companies to archive and make searchable all organizational email. At the same time that email is becoming more important, the volume of extraneous spam is burgeoning, threatening to overwhelm system administrators and users alike. Email management systems now must not only provide scalable, pinpoint search capabilities but also include features that are almost paradoxically opposite in nature, such as spam filtering, screening the dross from the gold — automatically, but very, very carefully.

Scalability is key to success in all of these varieties of search. But scalability alone is not sufficient: The total cost of ownership is also crucial. IDC believes that a well-architected approach can move search from the cost side of the ledger to potentially becoming an important driver of organizational profit.

When considering the total cost of ownership, ROI should be considered from a number of perspectives. One of the most important is whether search should be installed within the enterprise or outsourced. IDC has developed new models for hosted search that indicate an outsourced solution may be cost effective for many organizations, provided other factors such as security, accessibility, and network latency are adequately addressed.

Each of these categories needs to be considered separately in deciding on an appropriate search engine. In some cases, all the demands can be handled by one application, although it is possible that specialized applications for each may be more appropriate. Anyone implementing a search and retrieval system needs to consider all the uses to which it will be put. This document discusses the specialized demands that industrial strength search places on the search application.

The Elements of Industrial Strength Search

Four crucial elements form the framework of industrial strength search:

- Hardware and architectural requirements
- Content requirements
- The user experience
- Administration tools requirements

Let's examine each of these elements in more detail.

Hardware and Architectural Requirements

Industrial strength search must operate in real time, offering the freshest possible information in response to user queries. Put differently, this amounts to “solid state search,” using an index cache architecture rather than relying on slower mechanical memory to handle user requests.

To support real-time search, the index of available documents must be constantly updated, with subsecond latency from the time information is added to enterprise information repositories to the time it is included and accessible via search. In practice, this means that the index architecture combines real-time freshness with large data volumes and extreme query rates.

The system must support massive growth, providing unlimited scalability. Ideally, this would be plug-and-play scalability — just add more hardware — but the system should be optimized so capacity requirements won’t exceed existing hardware capabilities anytime soon.

Like all mission-critical real-time systems, industrial strength search should employ fault tolerance, automatic backup, and redundancy capable of handling system glitches or hardware crashes without missing a beat. For organizations that cannot afford or choose not to make investments in hardware with these capabilities, IDC believes that outsourcing search is an attractive, secure, and viable alternative.

Content Requirements

The world of closed-architecture, structured information systems has given way to the messy, unstructured world of multiple file formats and open-ended servers often operating with laissez-faire publishing policies. Industrial strength search must provide support for both structured and unstructured data, including a wide variety of document formats, ranging from simple HTML documents to Microsoft Office or PDF documents, as well as content residing in highly structured Oracle and DB2 databases.

Access to these formats must be transparent — the user doesn’t care where or how information is stored, so the search system should be capable of finding and delivering relevant content no matter where it resides within the enterprise, provided the user has appropriate permissions to access it.

Advanced linguistics should be provided to allow users to search by concept or meaning rather than by literal keywords. And given the multinational operations of many enterprises, industrial strength search must offer support for multiple languages, including Asian language and other Unicode byte sets.

The system should be customizable to the specific needs of the organization and individual functional groups. Administrative tools should offer the ability to provide business-focused tuning of relevance, hardware perfect answers to the most frequent user

queries, bypassing the standard relevance calculation procedures altogether.

The User Experience

Users demand high-quality, relevant search results, yet many are unable or unwilling to craft queries that adequately express their information need. Industrial strength search must be able to interpret needs from the scantest of clues, using a variety of techniques to “understand” what the user is looking for.

Query improvement and enhancement tools are the first step, expanding, disambiguating, or rewriting queries or probing the user for more information. Once the query is thoroughly understood, good relevance-ranking algorithms are applied against the corpus of enterprise information, eliminating noise and culling all but the best matches.

Results presentation can also aid the user. Documents can be clustered into topical categories that are conceptually related. Other subtle clues, indicating document language, origin, date, author, and so on, can further help the user select the best fit. Above all, industrial strength search must offer graceful alternatives when no best answers are found — never leaving the user facing a dead end.

Administration Tools Requirements

Industrial strength search must provide a suite of feedback tools to help system administrators. At a basic level, these tools should provide the capability to tailor search forms and result pages to align with the look and feel of the overall system. They should also provide tools to modify or adapt relevancy algorithms to meet the specific needs of the enterprise, boosting the relevance of some documents and decreasing others.

The tools should also help administrators analyze user interactions with the systems, providing reports of successful and unsuccessful queries, the most frequent queries, and so on, not only to help tune relevancy but also to identify potential holes in content that should be filled with information that satisfies the need.

FAST Search Features and Modules

The Fast Search and Transfer (FAST) enterprise platform is FAST Data Search 3.0, a modular and scalable solution that delivers real-time search and filtering capabilities. FAST is headquartered in Oslo, Norway, with U.S. headquarters in Wellesley, Massachusetts. The company is listed on the Oslo Stock Exchange and reported revenue of \$36.1 million for 2001, with 2Q02 revenue of \$10.4 million, for a year-on-year growth rate of 22% and a net profit of \$0.8 million.

The FAST Search Engine Technology

FAST Data Search is a modular system that consists of a base package of software and a number of add-on modules. The product

suite combines a real-time search engine, a real-time filter/alert engine, advanced linguistics, and numerous content access options, including support for more than 200 file formats and structured data from Oracle and DB2 databases. All documents are converted to a uniform XML format prior to indexing. FAST Data Search is available as a software or ASP solution.

The FAST Data Search platform consists of four basic sets of modules:

- **Data aggregation:** Able to aggregate data across Web servers, file servers, databases, and enterprise applications as well as an open API for pushing information into FAST Data Search
- **Content preprocessing:** Extensive modular framework for data processing, which includes a range of standard modules such as categorization, language detection and tokenization, file format conversion, and automatic phrase and concept extraction
- **Real-time search and filter engine:** Contains both a high-performance incremental indexing search engine and a real-time filter engine that can efficiently handle both structured (e.g., XML) and unstructured data as well as textual and numerical data
- **Front-end query and result set processing:** Accepts queries, analyzes the queries — including advanced query rewriting (e.g., correction for misspellings) — and routes them to the appropriate search node(s) and/or stores them as triggers for the real-time filter; offers various options for results set postprocessing before returning the result set

FAST Data Search is highly scalable across each of the four module sets described above: A small system may consist of a single server with no separate computer for the content processing and front-end modules. Larger systems may combine many search servers with a number of separate servers for data aggregation, content preprocessing, and front-end query analysis and result set processing to handle larger content indices or heavy query volume.

In addition to a modular, highly scalable system architecture, FAST Data Search offers powerful relevancy features supported by advanced linguistics and query capabilities. Four key processes work in tandem to assure the highest level of relevancy of search results:

- **Understand the content.** At the time of indexing, content is extensively preprocessed and analyzed. The system identifies proper names, reduces words to their base forms through lemmatization, identifies synonyms, extracts concepts from pages, and automatically categorizes each document based on its subject matter. Link analysis adds another layer of understanding for hypertext documents. This preprocessing pipeline is also user extensible.
- **Understand the query.** Query processing begins with the system detecting the user's preferred language. Phrasing and antiphrasing are performed — phrasing to assure that common

phrases are detected and processed as a phrase rather than individual keywords and antiphrasing to remove nonessential “noise” words from the query. Next, the query is spell checked against both standard and custom dictionaries and corrected if necessary. Queries with no hits are automatically modified, and finally, the query is analyzed to determine whether it is a general question, focuses on a problem, or is a highly specific and narrow request.

- **Selective matching.** Once the content and query have been thoroughly understood and refined, the system runs the query selectively against the appropriate index nodes. Content can be organized into specific, topic-centric collections. The system also offers extensive relevancy tuning (more on that below).
- **Present results in context.** On a results page, the query is displayed as the user entered it, or as the system reinterpreted it. Query terms are highlighted in results, and results are clustered and categorized. Furthermore, FAST Data Search offers various options for query-time clustering and other means of efficient result set navigation, such as suggested refinement terms *find similar*, *exclude similar*, and so on.

FAST Data Search offers numerous controls for tuning relevancy. System administrators can add or remove content from indices and can modify the relevancy scores for any document. Relevancy scores for a document can be either absolute or relative. The administration system also provides the capability to customize meta tags.

System administrators can define “virtual collections” for different groups or users. This allows the administrator to specify what part of the whole index will be searched and apply different relevancy indicators from collection to collection.

FAST Data Search Add-On Modules

In addition to the core modules, three add-on modules are available:

- **Real-time filter engine.** The real-time filter accepts high-rate data feeds; matches the data against stored queries, called “triggers”; and issues alerts that can be delivered through any messaging gateway including fax, email, and WAP. The incoming content streams are matched to user profiles in real time, and the same real-time content may be simultaneously indexed for search. The real-time filter also supports complex query terms that may span several information sources.
- **Real-time search.** The indexing latency (time between when a document is added to a system and when it is searchable) for the core FAST Data Search product is less than 10 minutes. Users requiring faster indexing can add the Real-Time Search module, which reduces maximum indexing latency to 90 seconds.

- **Real-time instant access.** For users with even greater demand for instantaneous indexing, this module provides subsecond indexing latency, using memory-mapped indices.

Case Studies

FAST has numerous customers using its FAST Data Search platform, including Broadvision, Chordiant, Dell, IBM, and Reed Elsevier. Three case studies illustrate the flexibility of the platform as a solution for diverse business needs.

Reuters

Reuters is the world's largest international news and television agency, publishing over 30,000 headlines daily in more than 26 languages. The company serves more than 560,000 users in over 52,000 locations worldwide.

Business Need

Reuters desired to provide its customers with an enhanced capacity to deliver critical business information to their own customers over the Internet. To accommodate the volume of news stories published each day, the company required a scalable service capable of handling drastic increases in the volume of unfiltered real-time news.

Requirements

- Handle Reuters content, brokered content, or any news feed
- Provide for both personalized push (alerts) and pull (full-text search)
- Ability to scale real-time alerting and filtering to hundreds of thousands and millions of subscribers
- Ensure subsecond indexing latency

FAST Data Search was implemented by Reuters News Distribution Service (NDS), deployed as part of its Market Data Systems and Internet Finance platform. Using the FAST Data Search real-time content filtering and matching capabilities, each news story is instantly searchable. Using preconfigured triggers for individual users, Reuters can single out individuals who want to see a specific news story and deliver it to any number of those users in less than a second.

FirstGov

FirstGov.gov is the official U.S. gateway to all government information and services, including the federal government and local and tribal governments. Its ambitious goal is to provide a single point of access to the diverse and varied information- and service-oriented Web sites maintained by all branches of the U.S. government.

Business Need

The U.S. government is the largest publisher in the world. The site required unified access to all government Web sites and services, with information dispersed across tens of millions of Web pages and in a wide variety of file formats and across a large number of servers and applications.

Requirements

- Index 50 million documents, with the capability to scale to 200 million documents over a five-year time frame
- Maintain an overall index freshness of less than one week for static documents as well as accommodate real-time data
- Index information from numerous, diverse repositories in different physical locations, including databases and dynamically generated content
- Handle multiple file formats, especially the PDF format, which the government uses to comply with the Paperwork Reduction Act of 1995

Under contract from the U.S. General Services Administration (GSA), AT&T worked as the systems integrator to provide an outsourced hosted search solution, deploying the FAST Data Search product. FAST Data Search either met or significantly exceeded all of the GSA's RFP requirements, including total cost of ownership (TCO). Future plans call for enhancing results to allow users to request search results to be displayed by category, subject, and agency.

Banca IMI

Banca d'Intermediazione Mobiliare IMI SpA (Banca IMI) is the Investment Bank of the Sanpaolo IMI Group, one of the 50 largest banks in the world. Banca IMI provides financial consultancy services and operations to raise risk capital and debt and trades on its own account and on behalf of third parties over a wide range of financial products in both the regulated and nonregulated markets.

Business Need

Banca IMI's multiterabytes of historical structured transaction data are stored on Oracle databases. Banca IMI also uses Datasim, a management application for brokerage firms to perform transactions. Increased user demand was causing time-consuming bottlenecks and required unacceptable levels of investment in additional hardware and licensing fees just to keep pace with user demands.

Requirements

- Offload the query and information retrieval processing from the Oracle database to a more cost-efficient hardware platform

- Decrease the time to run daily batch reports from up to 10 hours to subseconds
- Decrease the time to run ad hoc reporting from 1–2 weeks to subseconds

The FAST Data Search solution significantly lowered the total cost of ownership for Banca IMI. System architecture costs were reduced because implementing FAST Data Search was significantly less expensive than increasing the hardware, Oracle, and application licenses required to support increasing user demand. Database maintenance costs were reduced as the size of the database decreased. Finally, overall database administration was reduced, enabling end users to get the data they need via a Web browser instead of relying on a database administrator to run individual reports as well as freeing the database administrators to handle a greater capacity of unique requests.

Industrial Strength Search Is Here to Stay

Once-adequate intranet or basic site search tools are increasingly overwhelmed, not only by the massive quantities of information they must sift through but also by new formats and new user demands that far exceed capabilities considered adequate just a scant few years ago.

Raw horsepower and scalability, though important, are no longer adequate for most enterprises. FAST Data Search, with its advanced linguistics, tunable relevancy algorithms, and flexible, scalable architecture, offers the kind of customizable options many organizations now require in a search application. It's an example of what IDC calls industrial strength search, which is going to be increasingly vital to the health and success of companies operating in today's brutally competitive information economy.

Document #: 28146

Publication Date: October 2002

Published Under Services: Search and Retrieval Technologies;
Content Management and Retrieval Software
