Reference Code: TA001681CKM
Publication Date: June 2009
Author: Richard Edwards

Butler Group
a Datamonitor Company

TECHNOLOGY AUDIT

Google Search Appliance 6.0

Google

BUTLER GROUP VIEW

ABSTRACT

The Google Search Appliance (GSA) is an on-premises search solution suited to either Intranet use or an Internet-facing Web site. Within an Enterprise setting, the GSA is able to search across corporate portals and Web sites, file servers and Content Management systems, and a growing list of business applications. Recent upgrades have improved the GSA's scalability, provided for increased customisation, and enhanced search results. The GSA integrates well with enterprise Identity and Access Management solutions, and is thereby able to ensure strong, document-level security features. Over 200 file types can be indexed and searched, and native 'Content Connectors' are available for a range of popular Enterprise Applications. Federated search is now supported, but only across GSAs. The GSA is offered in two different models. The GB-9009 can handle 30 million documents in a small (5U) footprint, while the GB-7007 scales to 10 million documents. Complemented by the Google Mini, Google Site Search, and a new administrative API, the GSA is fast becoming a compelling Enterprise Search solution.

KEY FINDINGS

- ✓ An 'in-the-box' offering. Minimal installation and support overhead. Predictable costs.
- ✓ Similar look-and-feel to Google.com, or organisations can brand and style.
- ✓ Can be customised and tailored to suit particular users and groups.
- ✓ OneBox for Enterprise provides access to enterprise applications and content.
- ✓ Integrates with various Single Sign-On and authentication systems.
- ✗ Federated search is only available between GSAs.

Key: ✓ Product Strength ✗ Product Weakness ⓘ Point of Information

LOOK AHEAD

Butler Group expects that Google will extend the functionality of the GSA over the coming months by introducing more Content Connectors and partner offerings.

FUNCTIONALITY

The information required for business decision making and operational activities is contained within both structured and unstructured data sets, and these in turn are generally stored within siloed repositories and IT systems that are scattered across the enterprise. Furthermore, an increasing amount of actionable business information is now stored beyond the corporate firewall, either with partners and suppliers, or with customers and communities. Only through the application of Enterprise Search solutions can information workers ever expect to find all of the information required to complete a task and do their job; hence the continued 'arms race' amongst vendors competing in this very important arena.

The market for Enterprise Search solutions has skyrocketed in recent years, as businesses and institutions seek to reduce the amount of wasted time and effort that is often part of 'information work'. With over 70% of salaries now related to information work, one doesn't have to be an accountant to work out that, if spent wisely, an investment in Enterprise Search technology can offer significant benefits to the business.

2006 was the year when Enterprise Search moved onto the corporate agenda, and ever since then vendors have been clamouring to get a piece of this valuable market. Autonomy, Endeca, Exalead, Google, IBM, Microsoft, and Oracle are just a few of the vendors with offerings in the Enterprise Search space, and new entrants with ancillary and niche offerings are appearing all the time.

Product Analysis

Getting the right information, to the right person (or process), at the right time is essential if organisations want to be agile and flexible. In the past, Enterprise Search solutions were the domain of 'information experts', requiring not only skilled knowledge of how to use the tool to best effect in order to get the desired results, but also significant amounts of administrative support for configuration, as well as server power for running the indexing and search processes. First launched in 2002, the Google Search Appliance (GSA) provides a solution for those organisations looking for an alternative approach, which, for the most part, removes most of these issues.

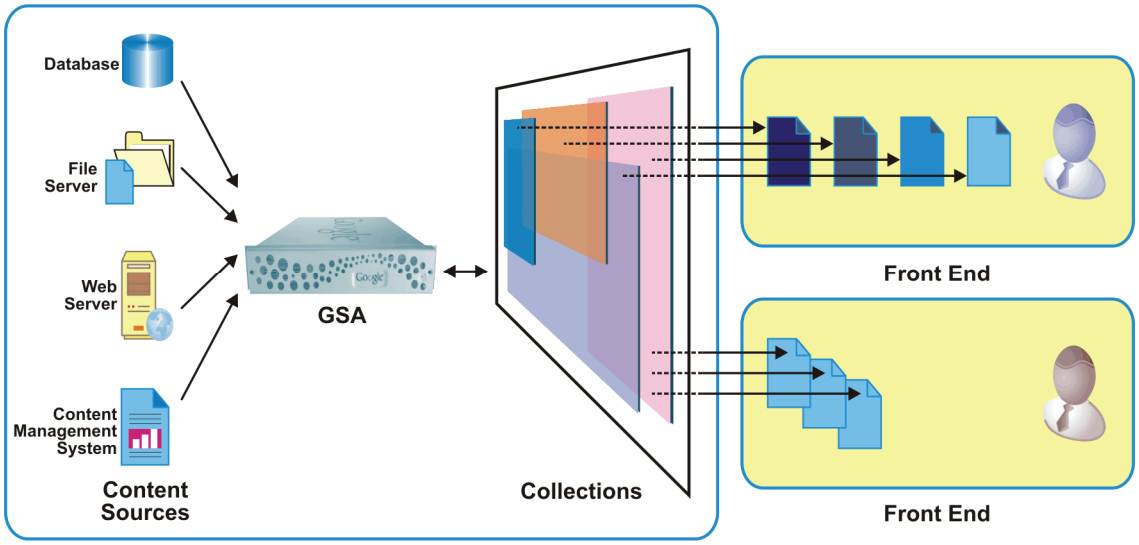
The GSA provides a 'universal search for business', in that most of an organisation's content can be searched via a single appliance. A single GSA can search Intranets, Web servers, corporate portals, file shares, databases, Content Management systems, line-of-business applications, Enterprise Applications, and Business Intelligence (BI) systems through the Google OneBox feature. Of interest to the many organisations with SharePoint installations, the GSA can index this content. Moreover, Google offers a free SharePoint WebPart so that users can access GSA search features from their SharePoint portal. The GSA can also index content stored in Lotus Domino. Using Google's simple XSLT style sheet, organisations can integrate the GSA into a corporate portal framework.

A hardware and software solution, GSA offers sophisticated Enterprise Search features without the complexity that is traditionally associated with this technology. Moreover, as the Google search interface will already be familiar to those likely to use the product, organisations adopting this solution are likely to experience rapid uptake by casual users, while 'Power Users' will also appreciate the more sophisticated facilities that are offered in the enterprise product.

The core technology of GSA is undoubtedly the ranking algorithms that enable users to find relevant content irrespective of source, location, or language. GSA can search across 220 file types, including HTML, Microsoft Office, PDF, PostScript, WordPerfect, and Lotus. With large amounts of legacy content stored in archives and proprietary repositories, organisations need to ensure that all information and data is available. GSA falls short of being an eDiscovery product, i.e. the kind of solution used to gather information in legal or regulatory proceeding, but its reach and range is impressive. The ability to search corporate e-mail accounts requires the user to install Google Desktop Search, and while this might suit some users with laptops or dedicated desktop PCs, it is not really a workable solution for those organisations where users share computing resources.

Users familiar with the Google Web search experience will notice some valuable facilities when they first use the GSA. First among these is the Dynamic Results Clustering facility. This enables the user to drill-down on a specific subject by automatically grouping search results by topic. Google recently introduced a similar facility on its Web search site called 'Wonder wheel'. Butler Group presumes that this feature, together with other recently introduced facilities, will be offered to GSA users in the near future. Organisations using controlled languages and taxonomies should assess the effectiveness of Google's algorithms in a pilot project, as this is one area in which human categorisation is often required to make sense of information. Google's algorithms make use of indexed content to infer spelling and other attributes, and so the company believes that that the GSA is well suited to almost all organisations.

Figure 1: Google Search Appliance Operation



Source: Google

DATAMONITOR

Unlike the Web, where metadata is relatively abundant and pages/documents are linked, the information stored within most organisations is at best 'fuzzy'. By this we mean that it is often imprecise, poorly 'labelled', and lacks any form of 'connectivity' with other documents or artefacts. Moreover, because the majority of documents are unlikely to be proofed, spelling mistakes and classification errors are to be expected. It is therefore important that the search indexing processes take this into account, and the search interfaces also accommodate mistyped search entries or phonetic searches for those words that may be incorrectly spelt. GSA includes a self-learning spell-checker, together with support for synonyms and stemming in multiple languages.

The GSA uses an algorithm to generate spelling suggestions from the content in its index. This means that it is especially good at providing relevant spelling suggestions for proper nouns, such as the names of employees and product names. However, because the spelling system is fully automated, it is not possible to manually edit the spelling dictionary. Synonyms can be used to suggest alternate searches for common spelling errors or company/industry-specific terms.

Recognising the fact that enterprise content is often devoid of metadata, the GSA is able to enrich unstructured data with metadata to allow filtering and customised search output. A recent addition to GSA is the ability to bias results based on metadata. Accessible only to system administrators, this feature builds on the ability to bias based on source, URL, or date. Google Suggest for Enterprise automatically offers suggestions to search queries based on user queries and enterprise content. The GSA can also be configured to allow user-added results.

GSA's e-mail alert feature enables users to subscribe to topics and documents of interest, and can also be used to populate corporate portals or online project workspaces via an API call or the e-mail route. Once again this feature is not intended to deliver a desktop contextual search service (whereby the search facility takes into account the documents being worked upon by the user), but it does offer a valuable feature that will be of use to many busy corporate employees. Alerts can be scheduled hourly, daily, or weekly.

The ability to customise and tailor the search experience to suit particular users and groups is an indication that Google understands the needs of large organisations. Indeed, Butler Group defines the phrase 'Enterprise Search' to mean any search product that can address all of the various needs and requirements of a business or institution. GSA's ability to search across file servers, relational databases, Microsoft SharePoint, and other Enterprise Content Management systems is another indication of this product's maturity. To index all content within a SharePoint site the Google Enterprise Connector for SharePoint must be installed on a host computer (Windows or Linux). This is required because SharePoint uses JavaScript, and the GSA does not execute JavaScript during the crawl and indexing process. It is also worth noting that Google's Connector Framework allows third-parties, including corporate developers, to develop new connectors should these be required. Google's Solutions Marketplace currently lists over 50 Connectors, eight OneBox modules, and six search extensions.

It should be noted that real-time information provided via the OneBox feature requires a separate server. The OneBox 'trigger' determines if the query is relevant to a given OneBox module. Triggers can be as simple as keywords or as sophisticated as regular expressions. Google OneBox for Enterprise allows real-time results to be displayed for many common enterprise applications, including Sales force, SAS, and others.

Other GSA features of note include the ability for administrators to view and export various important usage statistics including Google Analytics. The GSA offers a range of reporting features, and as the GSA tracks every user, query, and click, much can be gleaned. The GSA also offers drill-down into reports on pages that returned errors when the GSA tried to crawl them. The GSA enables systems administrators to set multiple user levels, from super administrator to collections manager. The collections manager is able to establish collections that users can search across, but is not allowed to change the settings for accessing secure content.

With GSA 6.0, Google has introduced an XML-based set of APIs to all administrative functions. Moreover, administrators can now manage a single appliance or a network of GSAs. Also new with this release is the ability to federate search amongst GSAs. Federation is an important feature for large, geographically distributed organisations, as it enables organisations to distribute the search technology close to the source content without losing the ability to perform enterprise-wide searches. Although this federated search feature is a welcome addition to the GSA, the fact that federation is limited to GSAs still presents a challenge to those organisations using other search solutions.

The GSA integrates with various Single Sign-On systems, such as Lightweight Directory Access Protocol (LDAP), NT LAN Manager (NTLM), and Windows Integrated Authentication. The search appliance also integrates with form-based Single Sign-On systems, such as those from Oblix, Netegrity, and Cafesoft. The GSA also provides native support for Kerberos. Google also has a Security Assertion Markup Language (SAML) Service Provider Interface (SPI) that can allow organisations to integrate with other security systems.

Product Operation

Internally, the functionality of the GSA is divided into three areas: locating content, indexing content, and serving search results. Content is located in three ways: Crawling, Traversal, and Feeds.

Crawl is the process by which the GSA locates content to be indexed. Crawl is a pull process, where the search appliance pulls content from the content location. The search appliance can also crawl a relational database to obtain metadata.

Traversal is the process by which the GSA locates content to be indexed in a content repository, such as EMC Documentum or Open Text Livelink. The Connector issues queries to the repository to retrieve document data to feed to the Google Search Appliance for indexing.

Feeding is the process by which content is directed to the GSA instead of having the search appliance locate content. Feeding is a push process, in which the content files are pushed to the GSA.

Indexing is the process of adding the content from the crawled documents to the index. After a file is retrieved by the crawl, the file is converted to an HTML file and submitted for indexing. The indexing process extracts the full text from each content file, breaks down the text, and adds both the text and information, such as date and calculated relevancy markers, to the index so that users' search requests can be satisfied. The index and the HTML versions of each indexed file are stored on the search appliance (access to the HTML cached versions remain security controlled and can be disabled).

Serving is the process by which the search requests from users are satisfied. A user types a search term into the search box and the request is transmitted to the serving software. The search appliance locates results in the index based on our relevancy algorithms. The search appliance then returns the most relevant results to the user's browser as a series of links. When the user clicks a link in the results, the content file is displayed.

During ordinary operation, all three processes are running concurrently. While the GSA is serving results it is also locating new content and indexing new or updated content.

Product Emphasis

Vendors and analysts often refer to 'out-of-the-box' functionality; however, the GSA's functionality is 'in-the-box' – designed to accommodate all but the most sophisticated search and discovery requirements, it brings the familiar and intuitive interface of Google.com into the enterprise environment. Besides the two appliances for larger organisations, Google also offers the 1U Google Mini, which will index from 50,000 to 300,000 documents, and as with the GSA, will provide up to 25 queries per second. Google's building-block approach to search means that organisations can increase the scale of their search infrastructure by simply adding more GSAs; however, with the introduction of the GB-9009, organisations can also opt for a single-box solution that is capable of scaling to 30 million documents.

DEPLOYMENT

The Google Search Appliance is delivered as a self-contained rack mounted 2U unit. The unit contains all of the dependencies and resources (for instance databases and storage etc) needed to support search. If required, redundancy and increased throughput can be achieved by deploying multiple appliances, and placing a load-balancer before them (see below). To index data using the custom feeds interface, a separate Java EE (JEE) server instance is required to support the Connector manager.

The Google Search Appliance comes in two models depending on the number of documents to be indexed:

- GB-7007: A rack-mounted two-unit (2U) appliance that can be licensed to search from 500,000 to 10 million documents. Multiple GB-7007s can be linked together to support larger document counts, or to integrate search across multiple departments, geographies, or Web sites. The GB-7007 replaces the GB-1001 for all new customers.
- GB-9009: Well suited for centralised deployments that support multiple business units or very large Web sites. The rack-mounted (5U) GB-9009 can search up to 30 million documents out-of-the-box. The GB-9009 replaces the GB-8008 for all new customers.

Custom search solutions can be built on request for larger deployments, with the GB-9009 capable of being scaled-out to support billions of documents. For any given document index size, the GSA has been tuned to provide a comparable user experience to google.com.

Each appliance has a level of fault tolerance built in. For example, although the GB-7007 is a single node unit, it has RAID architecture to provide tolerance against disk failures, and a second power supply unit built in. If additional levels of fault tolerance are required, multiple individual appliance units can be deployed in parallel. A single XML file configuration file describes the entire configuration state of a GSA. This can be version-controlled and exported to any additional backup units. The GSA is self-monitoring. If it encounters a problem it can alert administrators using standard protocols which can be integrated with popular system monitoring tools such as HP OpenView.

In comparison to more traditional Enterprise Search solutions, the GSA is very easy to install and configure. Once plugged into the corporate network, configuration is performed via an easy-to-use Web interface. For integration with existing data sources which have complex security models and/or can not be crawled directly, administrators will require knowledge on those data sources. A range of Google Enterprise Partners are available to assist in more complex deployments. Google states that two to three weeks is typical for straight-forward Intranet deployments. More complex deployments can take up to a few months where complex security models and/or source systems are involved.

Software updates are available from the Google Enterprise support site, and can be uploaded to the GSA using the built-in Version Manager component. These updates contain new features and improvements and are released typically in a six month cycle. Customers can choose if they wish to deploy the updates. The GSA provides a built-in testing framework for new updates, meaning that administrators can rollback or accept changes after testing it first.

Once running, the GSA is designed to need little or no maintenance. The self-learning algorithms mean that search quality actually improves over time with no intervention from an administrator. Any adjustment of the configuration (for instance to add new synonyms, run search reports, etc.) can typically be performed by existing personnel with a knowledge manager role. The Google Search Appliance comes with enterprise-level support included in the pricing for the duration of licence period. This support can be provided via telephone or e-mail, or via the Web. Additional support can be purchased if organisations require 24/7 or on-site support. If the GSA was sold by a partner, then they provide first-line technical support.

PRODUCT STRATEGY

Vendors and analysts often refer to 'out-of-the-box' functionality; however, the GSA's functionality is 'in-the-box' – designed to accommodate all but the most sophisticated search and discovery requirements. Alongside the two appliances for larger organisations, Google also offers the 1U Google Mini, which will index from 50,000 to 300,000 documents.

The GSA's target market is very much horizontal, and addresses the requirements of companies of all sizes and industries. The GSA can be used for internal Intranet searching or it can provide Internet searching of public-facing Web sites. Although the GSA is not aimed at any industry-specific vertical, Google's partners can offer domain expertise to support deployments in particular industries or integration with specific applications.

The licensing arrangements for GSA are deliberately simple and priced upon the number of documents to be indexed rather than the number of users or the physical volume of the data. This makes the product suited to either Intranets or Internet use. Entry cost for the GSA GB-7007 is US$30,000. This includes two years' support for both the software and hardware, with replacement of the appliance in case of failure. This also covers the transfer of the index from one GSA to a replacement if required.

COMPANY PROFILE

Google is a technology company focused on providing Web search and advertising. The company maintains a large index of Web sites and other online content, which is freely available through its search engine. The company generates revenue primarily by delivering online advertisements. Its major revenue sources are Google AdWords and AdSense. Google interface is available in more than 120 languages. The company operates in the US, the UK and a large number of other countries. Google offers a broad portfolio of Web-based products and services which are classified into six categories: Google.com, applications, client, Google GEO, Google Mobile and Android, Google Checkout, and Google Labs. The company primarily operates in the US, the UK and other countries. It is headquartered in Mountain View, California and employs about 20,222 people. The company recorded revenues of US$21,795.6 million during the financial year ended December 2008 (FY2008), an increase of 31.3% over 2007. The US, Google's largest geographical market, accounted for 48.8% of the total revenues in FY2008. Revenues from the US reached US$10,635.6 million in 2008, an increase of 22.3% over 2007. The UK accounted for 13.9% of the total revenues in FY2008. Revenues from the UK reached US$3,038.5 million in 2008, an increase of 20.1% over 2007. The Rest of the World accounted for 37.3% of the total revenues in FY2008. Revenues from Rest of the World reached US$8,121.5 million in 2008, an increase of 51.4% over 2007.

Table 1: Financial Details

Year ending 31 December	2008	2007	2006
Revenue (US$ Million)	21,795,550	16,593,986	6,138,560
Change on Previous Year (%)	31.3	56.5	72.8
Total Net Income (US$ Million)	4,226,858	4,203,720	3,077,446

Source: Google

DATAMONITOR

SUMMARY

Google's mission is to organise the world's information and make it universally accessible and useful. As a first step to fulfilling that mission, Google's founders Larry Page and Sergey Brin developed a new approach to online search that took root in a Stanford University dorm room and quickly spread to information seekers around the globe. The Google Search Appliance can natively crawl and index documents in popular Content Management systems, including EMC Documentum, IBM FileNet, Microsoft SharePoint, and Open Text Livelink. Furthermore, the Google Search Appliance has an open content connector framework that enables organisations to connect the appliance to virtually any other content management system. The GSA integrates well with existing security and access control systems, and offers end users many of the same benefits they have come to expect from Google.com, with specific enterprise enhancements that make search easy, useful, and intuitive.

Table 2: Contact Details

Google – London Sales & Engineering Office	**Google Inc.**
Belgrave House	1600 Amphitheatre Parkway
76 Buckingham Palace Road	Mountain View
London, SW1W 9TQ	CA 94043
UK	USA
Tel: +44 (0)207 7031 3000	Tel: +1 (650) 253 0000
Fax: +44 (0)207 7031 3001	Fax: +1 (650) 253 0001
www.google.com/enterprise	

Source: Google

DATAMONITOR

Headquarters
Shirethorn House,
37/43 Prospect Street,
Kingston upon Hull,
HU2 8PX, UK
Tel: +44 (0)1482 586149
Fax: +44 (0)1482 323577

Butler Direct Pty Ltd.
Level 46, Citigroup Building,
2 Park Street, Sydney,
NSW, 2000,
Australia
Tel: + 61 (02) 8705 6960
Fax: + 61 (02) 8705 6961

Butler Group
245 Fifth Avenue,
4th Floor, New York,
NY 10016,
USA
Tel: +1 212 652 5302
Fax: +1 212 202 4684