



An Oracle Technical White Paper

July 2013, Release 11.2.2.2

Oracle Secure Enterprise Search 11g R2

Content

Executive Overview	1
Introduction	1
The SES 11gR2 Release.....	4
Facet Navigation.....	4
Support for Multi-Tier Topology	7
Changes in the Technology Stack	8
Support for Search Result Tagging.....	8
New Sort Options for Search Result List ('Hard Sort').....	9
Push Crawling Mechanism	10
New Unified Microsoft Sharepoint Connector	11
Other New Features	12
SES Architecture	12
Crawler	14
Search UI	16
Administration.....	18
Search quality.....	20
Secure search	21
SES Methodology.....	22
The Gather Step.....	22
The Analyze Step	22
The Maintain Step	23
Robust Connector Framework.....	23
Security Plug-In Architecture	24
Performance and Scale	25
Concept Search and Result Clustering	26
The Clustering / Topic Interface.....	27
Powerful Search Query Syntax.....	28
Thesaurus & Alternate Query Terms	28
Attribute Shortcuts	29
Document Service Interface	30

Embedding SES as a Search Service.....	31
Other Features	32
Conclusion	32
Further Reading	33

Executive Overview

Secure Enterprise Search 11g (SES) is Oracle's search solution which securely covers all enterprise sources, and is easy to use and deploy. SES provides:

- Excellent search quality, going beyond keyword matching
- Sub-second query performance
- Scale and manageability for searching a large enterprise (100, 000 users, tens of TB of data, across a comprehensive set of secure content and application data sources) using commodity hardware
- Out-of-the-box user experience

Introduction

Internet searches have shown that significant information uplift can accrue from search technology. Without search engines, the Internet would still have billions of web pages, but surfers would have to know URLs *a priori*, or navigate through directories, to locate pages of interest. Clearly it is Search that makes Google popular, and the Internet more useful even as the amount of information on it grows at a rapid pace.

Proliferation of information also exists in the enterprise; however enterprises have so far not benefited from the information uplift that good search provides. This has been largely due to the differences between the Intranet and the Internet. For example:

- Information on the Internet consists overwhelmingly of web pages. In the Intranet, information – data and content – is spread across web pages, databases, mail servers or other collaboration software, document repositories, file servers, and desktops. An Intranet search engine must be able to search an organization's web content, its applications, databases, and mail through the same interface. Comprehensiveness,

When we look at the way organizations use information it is easy to see that information search is easily more than half the picture. While information creation (through data entry, document creation, writing emails etc.) is an unavoidable cost, information search is the primary mechanism for creating value from information. The fact that search has only recently become important demonstrates how little attention it has been given. As information sources proliferate so search will take on much greater importance.— **The Business Value of Enterprise Search, Report by Martin Butler Research, 2009**

across structured and unstructured sources, and ability to reach every corner (the *deep* Intranet) is the key to Intranet search.

- Unlike the Internet, where all information is publicly visible, Intranet information needs to be secure. Different users have different access rights to information, and information resources are often password protected. An Intranet search engine must be able to enforce security. If a user is not authorized to see a document, email message or record – then even the existence of the record should not be visible to him. The access rights can change, and access-changes made to the different underlying information-stores have to be propagated to the search-engine quickly.
- Internet search engines like Google use the links that URLs provide between web pages to deduce the importance or relevance of a document in a given search. Unfortunately, Intranet resources do not invariably vote for each other by URL links: a document authored in PDF may not link to the database record of a customer that it describes. Consequently, different techniques are needed for high relevance when it comes to Intranet search.
- Different Intranet users not only have different access-control rights to resources, but they also have different information needs based on job function. Search results have to be personalized to meet those needs.
- Higher service level expectations exist for the Intranet, and the robustness of an Intranet search product must match that of mission-critical enterprise software.
- Intranet search software must be simple to use and administer.

SES solves the problem of finding relevant information across your company's many disparate repositories of information, providing a very intuitive interface to search and administer.

The screenshot shows the Oracle Secure Enterprise Search 11gR2 interface. At the top, there are navigation links for "All", "Home Depot", and "Conference". A search bar contains the text "orac" with a "Search" button and "Attribute Filters" and "Browse" options. Below the search bar, a "Suggested Links" section lists "Oracle Database Homepage" and "Oracle Database". A "Content Spotighting" callout points to the "Oracle Database Homepage" link. An "Auto Suggestions" callout points to the search bar's dropdown menu, which shows "oracle database", "oracle secure enterprise search", and "oracle rac". A callout "... or browse all searchable content" points to the "Browse" button. On the right, an "Online Help" callout points to "Help", "Preferences", and "Login" links. Below the search bar, the results are displayed for the query "oracle". A "Facet Navigation" callout points to the "Refine Results" section on the left, which includes filters for "Session Category", "Session Type", "Session Date", and "Session Time". A "Flexible Sort Options" callout points to the "Group by" and "Sort by" dropdowns. A "Search hit customization" callout points to the first search result, "CON9052 - Oracle Fusion Applications: Oracle Exalogic and Oracle Exadata, the Ultimate Platform".

Figure: Example of the SES 11g search result page in action. Note the new facet navigation feature seen on the left hand side of the search screen.

Oracle's Secure Enterprise Search

Oracle has developed text and information retrieval technology for over 15 years. The base underlying capabilities of Oracle Text (a comprehensive API) have long been available with the Oracle Database. Oracle Ultra Search was introduced with Oracle9i to enable a portal search across different repositories, and was available with the Oracle Database, Application Server, and Collaboration Suite. Building on these products, Oracle's Secure Enterprise Search technology adds several key capabilities.

- **Simplicity.** A simple out-of-the-box web user interface, for both search and administration - that has both the clean look-and-feel and ease of use that users prefer on Internet searches.
- **Comprehensiveness.** The ability to search across all your sources – web pages, files in file servers or desktop drives, databases, applications, mail servers and groupware, and more.
- **Connectivity to Legacy Repositories.** SES allows companies to access their most valuable assets – information about its specific business, its processes, products, customers, and documents that previously resided in proprietary repositories. Connectors include interfaces for EMC Documentum, Microsoft SharePoint, IBM Lotus Notes, Oracle's E-Business Suite, Oracle Siebel, among others.
- **Security.** The ability to search password protected sources securely. Oracle's search technology provides single-sign-on (SSO) based security where available, and can also employ application-specific security where SSO is not available.
- **High quality search results.** Brings for the Intranet a high level of relevance that users associate with Internet searches.
- **Going beyond keywords.** As the volume of information grows, users need advanced search techniques like the ability to categorize and cluster search results for iterative navigation.

SES is fully globalized and can search in all major languages, including Western European, Chinese, Japanese, Korean and many more.

SES is robust and enterprise hardened. Many hundreds of queries per second can typically be served off a modest Linux machine. Typical enterprise Intranets typically run into terabytes, and Oracle's search infrastructure has been repeatedly deployed for multi-terabyte loads.

The SES 11g R2 Release

11.2.2.2 is the first release of the 11gR2 series, and the first major Oracle SES release since 2011. We are adding a number of new capabilities, including facet navigation, multi-tier install, search result tagging, global sorting of search results, push crawler, and a newly redeveloped connector to Microsoft Sharepoint that supports all versions up to, and including, Sharepoint 2010.

Facet Navigation

The most extensive new feature area in SES 11gR2 is facet navigation. New Administrator GUI screens make it easy to create and extend facet trees and nodes. Oracle SES requires no coding to work with facets, but it also provides full API support for experienced programmers

Facet navigation is secure. The facet values and facet counts are computed using only those documents that the search user is authorized to see.

Faceted navigation is the dynamic clustering of items or search results into categories that let users drill into search results (or even skip searching entirely) by any value in any field. Each ‘facet’ (elements for navigational purposes are named Facets) displayed also shows the number of search hits within the search that match that category. Users can then “drill down” by applying specific constraints to the search results. Facet navigation is also called faceted search, guided navigation and parametric search.

Facet navigation:

- Supports exploratory use cases, in contrast to known-item search. It helps users who need, or want to, learn about the search space as they execute the search process.
- Facets educate users about different ways to characterize items in a collection or web site
- You want to hint users at related content they might not have thought of looking for, but that could be of interest to them
- Helps clarify query ambiguity when used with keyword search
- Can help when your site has too much content for it to be displayed through fixed navigational structures, but you still want it to be navigable

Faceted search is most useful:

- When you already have a tree of categories and your content items are categorized in multiple of those categories.
- For exclusionary filtering of search results. For example, once I have a list of all search results for “red tee-shirts”, I want to hide everything that isn’t my shirt size.
- When you already have reliable meta data from which facets can be created (for example, database records which contain a mix of unstructured text fields and categories from which facets can be created)

Facet search adds little value in these cases:

- Known-item search. Users are better served by a search box-only approach to specify an item by name to locate it.
- When you have mostly unstructured text content with little, or no, reliable metadata attached to each document. Data quality can be a bottleneck and you need to have accurate, understandable facets that relate to users’ information needs. Offering users facets that are either unreliable or unrelated to their needs is worse than providing no facets at all.

SES also provides topic clustering, a mechanism which offers dynamic result clustering on top of a subset of results deemed most relevant by the SES relevancy algorithm. Topic clustering automatically extracts the most important recurring phrases and presents them in a cluster tree. Navigation on clusters is one at a time, unlike the new facet navigation feature where browsing on multiple facets is

supported. Facet search is also faster – extracting phrases from documents takes time. Integration between topic clustering and faceting (“Topic facet support”) is planned for a future release.

It is easier to understand what faceted search is through a few examples:

- **Facets are a tree of nodes**

In Oracle SES facets are created as a hierarchical tree of ‘nodes’ on top of search attributes which must have been defined by the administrator before and whose values are used as the facet values. The example below creates a simple facet on top of the “manufactures” attribute.

Figure showing example of simple facet creation in the Administration GUI

- **Facets have a type (number, string, date) and can have ranges**

Assume you have a field called price for the documents and you have that field faceted. You can configure SES to return the facets as ranges of values (e.g., 0-\$100, \$100-500, \$500-1000, etc.) as shown in the example below.

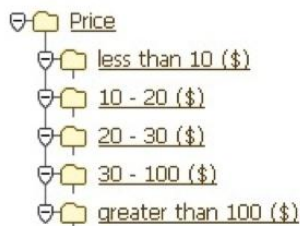


Figure showing creation of a facet node with ranges

- **Facet values can be flat or hierarchical**

Hierarchical facets allow for a structured categorization scheme of content items. End users interact with the hierarchy via drill down. When you click on the ‘kitchen (2)’ facet in the below example, a new facet is introduced - that of ‘sinks (2)’ as seen below.



Figure demonstrating how users interact with hierarchical facets in the SES Search UI

- **Multi-word searches can match multiple facets**

For example, “computer parts” may match “computer” facet and “part” facet.

Facets are first created either in the SES administrator interface or per Administration API.

Administrators can decide which facets are deployed for which source group.

The values of facets are collected during crawling and are available immediately when a crawl completes and end users can then start navigating facets via SES Query API. Facet definitions can be modified after crawling - there is only a small latency for introducing new nodes, for example. The values of document attributes are evaluated and matched against the facets, their nodes, and ranges. It is possible that a document is mapped to multiple facets

Support for Multi-Tier Topology

Oracle recognizes the need to give its customers the power to deploy SES more flexibly. Previously, the SES software stack could only be installed in ‘appliance mode’ -- standalone and as a whole. The SES installer would transparently install a new, separate Oracle database, a WebLogic Application server, and then deploy the SES application inside. The philosophy was to make installation as simple as possible and require as little advanced database or application server skills as possible. Installing SES into existing Oracle databases or Fusion Middleware deployments was not possible. With 11gR2, this way of installing the whole SES software stack ‘in one go’ is still available and we expect it will continue to be popular.

SES now supports three install options:

- Software Appliance -- Same as in earlier releases. Standalone install of the whole SES software stack including database backend, application server mid tier, and SES application itself.
- Existing Database -- Customers can use an existing Oracle database for the SES search engine index and data tables, while the SES installer creates a WebLogic Application Server instance to run the

SES application. The database can be single instance or RAC and must be 11.2.0.2 or 11.2.0.3 release level.

- SES software only -- Install SES software into both preexisting Oracle 11.2.0.3.x database and preexisting WebLogic Application Server. WebLogic can be single node or cluster and must be 10.3.6 WebLogic/Fusion Middleware 11.1.1.6 release level.

The new install options bring several benefits:

- Deployments of SES are now more flexible and require less footprint since SES can an Oracle database with other products
- SES can now be installed into existing multi-tier deployments
- Businesses with a need to expose search to both private network users (intranet) and external users (internet) without security compromise can install SES into an existing DMZ environment. This is easier with the new release, it requires less footprint. For example, one SES instance can be designated to serve the internet by placing its Application server component into a separate subnet with other hosts (DMZ) in order to be less vulnerable to attacks from users outside of the local area network.

Changes in the Technology Stack

In previous versions, the SES crawler ran in the embedded Oracle database and made use of database scheduling. With 11gR2, the crawler has been moved to the Application Server middle tier and uses Enterprise Scheduler Service (ESS), which is used to spawn crawl jobs.

Crawler schedules and frequencies are still configured via the SES Admin GUI, but Oracle Enterprise Manager can be used to pause or cancel crawling jobs and to set blackout windows. This provides for better scheduling capabilities.

Running crawl jobs from the application tier recognizes customers need to better secure their computing infrastructure. Customers often want to place a firewall between application server and database to restrict access to their databases to a known network route. In this scenario they typically don't want any network connection to originate from the database tier. The new architecture provides this.

Support for Search Result Tagging

With a new tagging feature, SES 11gR2 provides search end users with the power to influence search result rankings directly. End users can log into the search UI and enter tags for a search result of their choice. A match for a tag makes a difference in relevancy – search requests that match tag text will have their relevancy scores boosted.

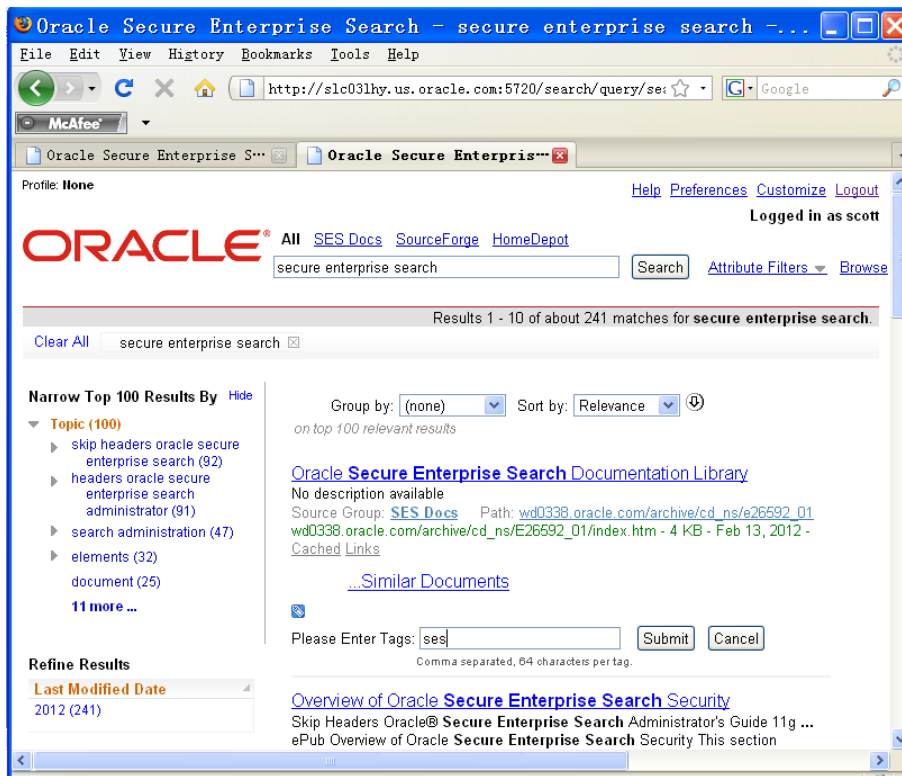


Figure illustrating the tagging of a search result with tag “SES”. Capturing user generated terms as tags to documents can improve relevancy scores and make documents more findable.

New Sort Options for Search Result List (‘Hard Sort’)

A new UI drop-down feature offers ‘global’ re-sorting of the search results by user selectable attribute. ‘Global’ means that sorting now extends over all matches of a given search term. In prior SES versions, only the top-N search results deemed most relevant by the relevancy ranking algorithm of SES would be re-sorted by attribute.

The administrator can configure which attributes are available in the drop-down list for sorting and choose a default ordering. In this way, end users are given a choice of sort options, and in some cases they will prefer ‘global’ over ‘top-N most relevant’ because they can yield different results. For example, take a search request for ‘oracle’ where the user wants the search results sorted by creation date (recency) of the document.

- A ‘global’ sort on recency will always return the newest documents containing the term ‘oracle’, even if these documents are not necessarily most relevant
- A ‘top-N most relevant’ sort on recency will only consider the top-N highest scoring documents, and return them sorted by recency. If the oldest documents happen to contain ‘oracle’ a large number of times, then these documents are likely to be deemed the most relevant by the SES relevancy

algorithm and will come back with the highest score. Suppose the newest documents contain the term 'oracle' only a few number of times, then these documents are likely to be deemed less relevant by SES. Now if the newest documents in the collection happen not to be in the top-N, they will not be presented to the user, even though the user specified a sort on recency.

Sort by multiple attributes is also supported.

Push Crawling Mechanism

There are now two general strategies for crawling data sources, 'pull-' and 'push-', which allow to crawl/extract data by accessing external sources/system. Previous releases of SES were limited to pull-crawlers which have the disadvantage that the index might not be accurate, since SES only learns about data changes when a crawl is executed. Index accuracy depends on the frequency of crawls and it is hard for the crawling to keep up with applications who have continuously occurring transactions. "Push"-crawlers are implemented on the side of the client system which includes the data to be indexed. They proactively update the index and thus need to have a direct (network) connection to SES. Accordingly, push-crawlers are more difficult to implement, but can reflect changes in the data more rapidly in the SES search engine index.

Push crawlers work as follows:

- A data source is registered with Oracle SES much like it is registered with the 'pull' crawler
- Sources send their documents to Oracle SES as soon as they are generated. Documents are submitted by posting them to a HTTP/HTTPS protocol endpoint via HTTP POST
- Documents need to be packaged via a 'feed' mechanism based on the 'RSS' standard (RSS are a family of web feed formats used to publish frequently updated works). The RSS format requires that the contents of a repository be exposed as XML documents. An example of an RSS feed is shown below.
- Feeds are parsed and their documents are indexed immediately when received and become available in near real-time. However, SES looks at the timestamp in a feed and if the feed already exists in its search engine index with a newer timestamp, SES skips indexing it.
- Error logs for the feed can be obtained by doing an HTTP GET at the push endpoint URL
- File attachments to documents are also supported. However, applications who want to present document content as attachments rather than inline data need to open up themselves to an external connection such that attachments can be fetched. Otherwise, the push crawler obviates the need for external connections.

```

<rss version="2.0"
<channel>
  <title>Contacts</title>
  <link>http://my.company.com/main.html </link>
  <description>This is an example of an RSS feed containing contacts</description>
  <lastBuildDate>2013-04-03T12:20:20.00Z</lastBuildDate>
  <item>
    <title>Example entry</title>
    <link>http://my.company.com/contacts?id=paul</link>
    <itemDesc xmlns="http://xmlns.oracle.com/orass" operation="insert">
      <documentdata>
        <author>Administrator</author>
        <accessURL>http://foo.com</accessURL>
        <lastModifiedDate>2012-12-12T12:22:22.00PDT</lastModifiedDate>
        <keywords>Content Contact</keywords>
        <summary>This is the summary of the document.</summary>
        <docAttr name="organization">Reports</docAttr>
        <docAttr name="country">Germany</docAttr>
      </documentdata>
      <documentAcl securityAttr name="EMPLOYEE_ID">OR9NH</securityAttr></documentAcl>
      <documentInfo><status>STATUS_OK_FOR_INDEX</status></documentInfo>
      <documentContent>
        <contentLink contentType="text/html"> http://my.company.com/reports.html</contentLink>
        <content contentType="text/plain">Paul Robinson, A240, Westland Drive</content>
      </documentContent>
    </itemDesc>
  </item>
  <item>
    .....
  </item>
</channel>
</rss>

```

'Channel' is another term for 'feed'

At the heart of RSS feed files are 'items'. Items describe documents you'd like to push into Oracle SES to be indexed

Link to an attachment

Actual content of the document specified inline. A link to an external content file can also be given

Figure: Example of a 'push' data feed containing two documents with employee contact information

New Unified Microsoft Sharepoint Connector

Oracle has unified Sharepoint connectors into one single plug-in, which supports legacy Sharepoint versions 2003, 2007, as well as the new 2010 version. The new plug in adds support for:

- HTTPS-enabled Sharepoint sites

- The new Sharepoint 2010 document rating feature
- Objects new to Sharepoint 2010: Records- and Asset library, blogs, pages, and team sites

Other New Features

- Search Suggestions “as you type” are now officially part of the product
- Ability to crawl “sitemap” definitions

SES Architecture

Oracle’s Secure Enterprise Search is a standalone, self-contained server for search; it operates as a “black box” that indexes information from the crawler and serves up the results. It comes with its own user-interface and administration; it does not, for example, need you to program using SQL or administer as a DBA.

Architecturally, as presented in Figure 3, the product is made up of five distinct components:

- **Crawler.** The SES Crawler is a Java process activated by your Oracle server according to a set schedule. When activated, the crawler spawns a configurable number of processor threads that fetch documents from various data sources. The crawler maps link relationships and analyzes them to avoid going in circles and taking wrong turns. Whenever the crawler encounters embedded, non-HTML documents during the crawling it uses filters to automatically detect the document type and to filter and index the document.
- **Database.** An Oracle11g database contains the SES-repository, which stores information about the repositories indexed by SES and the search engine ‘index’ (information collected by the crawler, filtered and indexed by Oracle Text).
- **Search UI & API.** SES provides a customizable out-of-the-box user interface to the Server. It also provides a web services API for building custom applications for querying indexed data, and contains interfaces for Basic Search Form, Advanced Search Form, Query Result Display, Help Page, Feedback Page, URL registration, and so on.
- **Administration Tool and Interface.** The SES administration tool is a browser-based application that you use to configure and schedule the crawler, configure the server, run several reporting features, and other similar tasks.
- **Federator.** SES also provides the ability to federate queries to other engines that implement their own search – mail servers, Internet search engines, and specific applications. These results can be combined and displayed together along with those results served off the internal index of SES Server.

SES is based on an Oracle11g database and WebLogic Server J2EE container which implements a web server to serve up HTTP. During installation, the SES WebLogic Server ‘application’—consisting of search and administration environments -- is deployed within this J2EE runtime environment. The Oracle database is custom built – configured and adapted for the special needs of a search engine.

Both SES database and web server are installed on the same machine. Moving SES database and search/ administration applications to different machines can be done in principle, but is not officially supported by Oracle today.

SES connects to Oracle’s SSO-infrastructure (OID) without the need for any customization – only simple connection parameters to Oracle Internet Directory (OID) must be specified.

The SES WebLogic Server application is connected to the database via JDBC -- the connection is defined by the following files:

- *listener.ora, tnsnames.ora, sqlnet.ora* (Oracle Net configuration files)
Path: `$ORACLE_BASE/seshome/network/admin`
- *data-sources.xml* (defines database connection of Administration application)
- *search.properties* (defined database connection of Search application)
Path: `$ORACLE_BASE/seshome/webapp/config`

The above files are automatically configured during installation and should be left unchanged. If other Oracle software is found during install, the SES installer creates a new listener with its own network configuration/port.

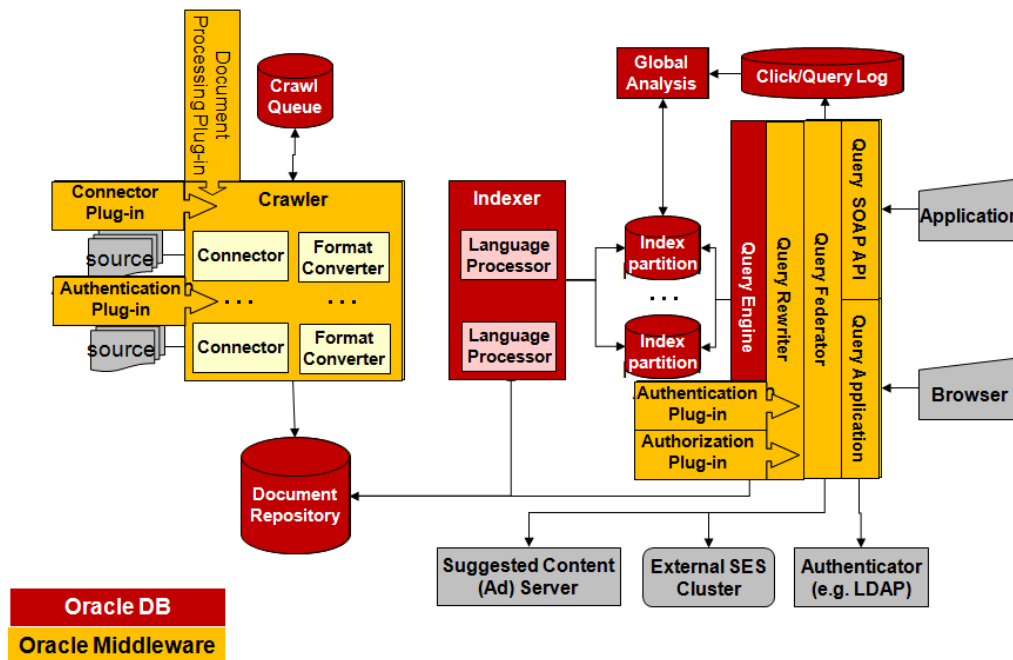


Figure Process-view of the Oracle SES 11g Architecture, including database- and web server (mid tier)-components

The following sections describe some of the components and other aspects in more detail.

Crawler

The SES crawler is a multi-threaded Java application responsible for gathering documents from the data sources you specify during configuration. To crawl different repositories, the SES crawler allows you to define specific 'data sources'. A data source is a logical construct identifying a repository. You can take a single physical repository, such as a database, and map it to multiple data sources (A data source is also the granularity at which you define metadata). SES knows a number of standard types of data sources (more data sources are available via connectors):

- Web Sites – Define web sites as a data source with the HTTP protocol like www.oracle.com.
- Database Tables – SES can crawl Oracle databases and other relational databases that support the ODBC/JDBC standard. Database tables to be crawled can reside in SES's own database instance or they can be part of a remote, database accessed over a network. SES allows the crawling of both full text columns and "fielded text" columns. Fielded text columns allow you to map a database column to an SES attribute (e.g. author, title), creating a set of indexes tuned to the content of your database.
- Files – Files must be directly accessible by the crawling machine. Remote files may be crawled so long as they can be crawled through the `file://` protocol. Files must be accessible by each crawler machine either locally or remote over the network.
- Emails. SES can connect to an IMAP email server and index all the emails for a user. To index mailing lists, you might choose to create a specific IMAP account, which is subscribed to the mailing lists of interest.
- Oracle AS Portal instances

SES uses 3rd party filters to extract text and metadata from documents and automatically identifies document types. The filters handle popular document formats like PDF and MS-Office. There is also support for filtering documents that have been compressed with ZIP utilities.

To maintain fresh, comprehensive search results, SES uses synchronization schedules. Email search results, for example, can be updated continuously, while published content is gathered on a less frequent schedule. Each synchronization schedule can have one or more data sources attached to it.

To limit the crawling to a specific section of your corporate network or to ensure that crawling does not take wrong turns and follow link relationships that point outside your Intranet, SES lets you specify so-called 'inclusion' and 'exclusion' domains for crawls. SES supports 'instance snapshots' where you create a read-only snapshot of a master SES instance for query processing or backup purposes. This is useful when the master instance is corrupted and you want to use a snapshot as a new master instance.

The SES crawler can be instructed to collect URLs without indexing them. This data harvesting mode allows you to examine document URLs and their status, remove unwanted documents, and start indexing.

Connector API

The crawler can be extended through ‘connectors’ (Connectors are Java classes supplied by Oracle, or developed by the customer, which run in the same J2EE container as the search application). Out-of-the-box, SES ships a family of connectors to a number of enterprise content sources like EMC Documentum, Lotus Notes, and Microsoft Sharepoint among others.

Connectors are deployed via the SES Administration GUI – they are listed as new data sources after being defined. After configuration, connectors supply data to the crawler and can be indexed just like other data sources.

Technically speaking, connectors are responsible for collecting URLs pointing to the documents to be indexed – they pass these URLs to the crawler for indexing. The following is a short sample of methods customers must provide if they want to implement their own connectors:

- *open*: initialize
- *startCrawling*: do any setup necessary for fetching documents
- *stopCrawling*: do not send more documents
- *isDeltaCrawlingCapable*: specifies whether the agent can return only documents created since a certain date
- *fetch*: return the next document URL
- *received*: acknowledgement that an URL has been fetched
- *getCredential*: return any user name and password necessary to access this URL
- *getCookies*: return the cookie stream needed to access the URL
- *getAttributeLOV*: return a List Of Values for all source attributes
- *close*: shutdown and cleanup

Web services API

Search engines are usually integrated in existing customer web- and portal sites. Ideally, end users invoke searches from a search mask and don’t even realize that Oracle SES handles their search requests in the background. “Look and feel” of the result list must correspond with the Portal site, which invokes the search. To achieve this, SES provides a web service interface, based on standards like SOAP and WSDL.

In the code example shown in the figure below, the end user enters their search term into an input field (“search box”). The search request is sent from the CMS application server directly to the SES web service. SES executes the search and sends results back to the calling application in the form of XML via SOAP. Results are displayed in embedded fashion – within the application.

SES uses no UDDI-repository --the WSDL-description can directly be obtained from the server.

```

import oracle.soap.transport.http.OracleSOAPHTTPConnection;
import oracle.soap.encoding.soapenc.EncUtils;
import oracle.search.query.webservice.client.*;

public class TestWS
{
    public static void main (String[] argv)
    {
        try
        {
            oracleSearchService search = new oracleSearchService();

            // Add your own code here, for example to populate
            // the query string.

            // Set SOAP URL. The URL is http://<host>:<port>/search/query/oracleSearch
            stub.setSoapURL("http://oes-serv-example:7777/search/query/OracleSearch");

            String queryString = "oracle";

            //
            // Do a simple search for the queryString we set up above
            //
            OracleSearchResult result = stub.doOracleSimpleSearch(
                queryString, // query
                new Integer(1), // startIndex
                new Integer(3), // docsRequested
                Boolean.FALSE, // dupRemoved
                Boolean.FALSE, // dupMarked
                Boolean.TRUE); // returnCount

            // Get the result set
            ResultElement[] resElemArray = result.getResultElements();
            // Loop through the results displaying the document title
            for (int i=0; i<resElemArray.length; i++)
            {
                System.out.println("Document Title:
                                   "+resElemArray[i].getTitle());
            }
        }
        catch(Exception ex)
        {
            ex.printStackTrace();
        }
    }
}

```

Typical code example of a search request via web service API

Search UI

SES incorporates the Freemarker templating engine, a Java library which makes customization of the Search UI very easy. The idea behind Freemarker is that you separate UI design from the actual program code. This allows for changing the appearance of a UI page without the need for changing or recompiling code, because the application logic (the SES Java programs) and the query page design (Freemarker templates) are separated. Templates do not become polluted with complex program fragments. Figure 2 below shows how a simple template processed by Freemarker will produce an “Output” page. Variables like “name” come from outside the template, and thus the template author has to deal with presentation issues only.



Figure illustrating the workings of the Freemarker template engine

In SES 11, Freemarker allows easy customizations of the SES default query UI, including:

- Changing the logo
- Changing the look and feel (colors, fonts)
- Modifying the page header and footer, including static text and links to other pages.

Freemarker allows for swapping in different designs, or ‘skins’, for the SES Search GUI. Individual skins are selectable through a URL parameter.

A skin is defined by its own sets of templates, images, CSS styles, JavaScripts, etc. Most important are templates. SES provides template files, written in Freemarker Template language (.ftl), which can easily be modified and customized. Figure 3 below shows how easy ‘results.ftl’, a Freemarker template defining the normal SES results page, can be customized. Here, changing the word “left” (highlighted) to “right” will have the effect of moving the “Filter Results By” sidebar of Figure 1 from the left side of the page to the right side.

SES provides a set of user interface components that can be imported as a library (seslib10.ftl) in the customer-modifiable templates. This separates UI “code” that is tightly coupled with internal SES logic (which may change between SES releases) with presentation code that can be freely edited to modify text copy and layout.

```

xtern
<!--results.ftl-->
<#import "/lib/oracle.com/seslib10.ftl" as ses>
<#compress>

<#assign searchFormName = "searchForm">

<!-- Set this to either "left" or "right" to position the sidebar on -->
<!-- either side of the page. -->
<#assign sidebarPageAlign = "left">

<#assign baseDocClass = "sidebar-" + sidebarPageAlign>
<#if showSidebar>
  <#if sidebarPageAlign == "left">
    <#assign docYuiClass = "yui-t2 ses-t210">
  <#else>
    <#assign docYuiClass = "yui-t4 ses-t210">
  </#if>
<#else>
  <#assign docYuiClass = "yui-t7">
</#if>

```

Figure showing xample of a Freemarker search result page template definition in SES 11

These components are implemented using user-defined directives, or macros, in Freemarker:

- Query box area, including source group tabs, input box, Search button, Advanced and Browse links.
Three formats: for splash page, top of regular page, and bottom of page
- Suggested links and Suggested Content tabbed display
- Collapsible sidebar
- Result clustering trees
- Result list
- Sorting and grouping drop-down lists
- Breadcrumbs
- Query stats (“Results 1 - 10 of about 28 matches..”)
- Pagination links (“Results page 1 2 3 4 5 ... Next”)

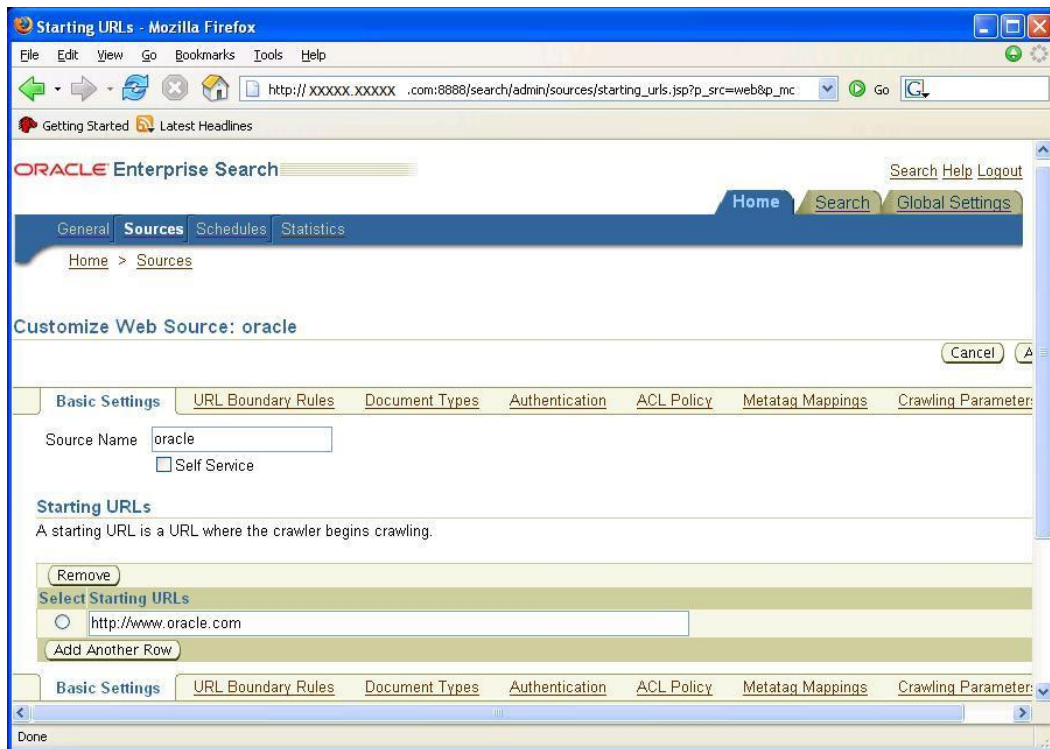
Administration

The administration tool is a web application that allows the administrator to:

- Define and crawl data sources.
- Define crawler parameters like URL boundary rules, crawling depth, language and proxy settings, etc.
- Create and modify schedules for the crawler.
- Set query options - Query options allow users to limit their searches. Searches can be limited to document attributes (e.g. title, author) and data groups. Data source groups are logical entities

exposed to the search engine user. When entering a query, the search engine user is asked to select one or more data groups to search from. Each data group consists of one or more data sources.

- Adjust relevancy ranking of the search hit list – SES allows administrators to influence the order that documents are ranked in the search hit list. Use this to promote important documents to higher scores and make them easier to find.
- Define suggested links for specific search terms.
- Define alternative words for specific search terms.
- Setup authentication mechanisms for certain data sources.
- Manage the backup and recovery of search metadata.



A typical configuration page from the Oracle SES Administrator GUI.

Administration API

An Administration API supports the management of large-scale deployments by providing a command-line interface to administrative tasks previously only available through the SES Admin GUI:

- Create, change, or delete sources or schedules
- Start and stop schedules

- Configure SES crawlers
- Failed operations are automatically rolled back

It even provides additional functionality that is not yet available in the GUI.

You can use the Administration API within an interactive session, or by executing commands from operating system prompt. A comprehensive set of help pages is available to assist with the command syntax.

Search quality

Search quality or the ability to find relevant information is one of the most important features of any search engine. SES uses a wide range of techniques for providing excellent search quality.

The following techniques are used at different stages of the crawling and index processes for enhancing the overall search quality:

- **Metadata processing.** It is very important to identify metadata from pages and documents like title, author, description, headline, email, and anchor text.
- **Duplicate elimination.** There is a lot content in a corporate Intranet that is duplication. Copies of same presentations, web pages, text documents are all over the place. Sometimes people produce multiple files and sometimes the servers duplicate the content for mirroring. Other issues with duplication are different versions, formats, HTML style, site-specific links, contact information, etc. In any case, the user should see only one copy of the document or web page when searching.
- **Complete duplicate elimination** helps to identify and remove duplicates at the crawling stage before the document is even indexed.
- **Link analysis.** One of the most widely used techniques for improving relevancy is link analysis. Briefly, the idea is to discover authoritative pages by performing analysis on the link structure of the web collection. A page that is linked by many pages is important. A page that is linked by a high link score page is also important. A number of algorithms exist today like HITS and PageRank. SES has its own algorithm implementation.

The administrator can also control the relevancy using a couple of extra features: alternative links and suggested words. Alternative links is a useful feature for registering a well-known authoritative page against keywords. These links will then be displayed at the top of the search result page when the user searches for these keywords. Suggested words can map user search terms to synonyms. For example, cellular phones for cell phones or wireless phones.

In case users have trouble spelling query terms, the spell checker feature suggests corrections based on data available from a dictionary and crawled data.

Apart from all the searching features of SES, it is possible to combine browsing and searching at the same time. You can click on the browse link to navigate all the directories that SES has created

automatically after the crawl. This is a good entry point when you are trying to explore all the content that is available to search. Of course, you can search within a directory at any time in the search box.

Secure search

SES features secure searching – the ability for users to log in and find documents which are not publicly available. To do this, SES has secure crawling capabilities, and the capability to store Access Control List (ACL) information alongside data sources

SES integrates with a number of Lightweight Directory Access Protocol (LDAP) servers such as Oracle Internet Directory or Microsoft Active Directory. These directories provide authentication ("who am I?") and authorization ("what can I do?") support to SES. Additionally, SES can use native authentication services for various sources (such as content management systems), which handle their own user databases rather than using an LDAP directory.

Authorization is handled by one of two models. In identity-based authorization, documents are tagged with a list of users and groups who have access to those documents. At query time, the authentication (identity) manager is responsible for returning a list of groups of which the currently-logged-on user is a member. Thus a query can be created which restricts the documents returned to those visible to the user (explicitly) or groups of which he is a member.

In attribute-based authentication, the source is responsible for defining a set (one or more) of security attributes for the documents it provides. A separate authentication plug-in provides the list of attributes which apply to a particular user. For example a source might define ROLE and RESPONSIBILITY as two security attributes. A particular document might have ROLE attribute values of "MANAGER" and "SALES". The authorization plug-in might return the information that "John Smith" has a ROLE value of SALES, which is a match against the document, and thus he is able to fetch that document.

The crawler can handle secure sources in a variety of ways. While the sources may themselves be protected by the same (or a different) identity server, this is not a requirement. Any source can be crawled in a secure manner, so long as it is protected by one of the following:

1. Oracle Single-Signon Authentication
2. HTTP Basic Authentication
3. Form-based Authentication
4. Service-to-Service Authentication (a trust relationship between the source and SES).

There are several different ways of providing access credentials to a secure crawler:

Admin Based Authentication

When a data source is defined, the administrator can enter an authorized password (either a user password or "super user" password). This will be used to collect the information from the source. An ACL may be defined for the source that defines who can search the information.

Self Service Authentication

When a data source is defined, the administrator sets up the source but does not provide any username or password. Users are then able to log in and provide their own access credentials. A data source is then created which is specific to that user, and they are the only user who can search that information.

Custom Agent or Crawler

A custom agent is a Java module that can be used to crawl any user-specified data. The agent passes back a pointer (URL) to the information to be indexed, and optionally specifies an ACL for each document. This allows great flexibility in access control.

SES also provides a QTA (Query Time Authentication) API that allows customer to have fine control on search results at query time. SES uses this technique as the main interface to filter documents based on authentication access.

SES Methodology

What steps do you need to follow for using SES? The SES search engine follows four logical steps to provide universal search – gather, analyze, make queryable, and maintain. These steps are not novel, and are indeed found in most organizations’ business process.

The Gather Step

Gathering refers to information that exists in structured relational databases and in unstructured files, Word processing documents, spreadsheets, presentations, e-mail, news feeds, Adobe Acrobat files, and Web pages. SES gathers this information by “crawling” your corporate Intranet and looking through all the information that exists in the various repositories of your company – databases, Web pages, IMAP mail servers and others.

During the gathering process, link relationships are analyzed to avoid going in circles and taking wrong turns. As a result, SES administrators have an easier time keeping search results complete and up-to-date.

The Analyze Step

In the analyze phase SES looks at the meaning and structure of gathered information. In order for information to be searched, it must be indexed. During the analyze phase, SES uses the Oracle Text engine to extract both meaning and structure from the gathered information by creating an integrated index, effectively “normalizing” both structured and unstructured data. Oracle Text indexes contain a complete wordlist along with other information.

During indexing, text and metadata are extracted from documents by third party filtering software. This filtering technology automatically identifies document type, invokes the correct filter and produces indexable text and data. Several predefined metadata fields are supported, including author, date, and title. The filters include the most popular file types like MS Office and PDF.

Unlike some document management systems, SES gathering and analyzing is non-intrusive. Instead of physically moving documents, information and documents are analyzed but reside in their original location under their own name.

In typical Web search technologies, hundreds of hits are returned. As the number of repositories increase, the ability to rank relevance of documents decreases. SES uses the award winning relevance ranking of Oracle Text to ensure that users consistently find the needle in the haystack.

Making crawling results searchable

“Make Searchable” is the function of providing access to all the information that has been indexed in a programmatic fashion. Oracle SES provides a web services API for this purpose. Passing a search term into the query API locates all relevant documents, whether they are stored on Web servers, databases, or in applications. Customers can use SES APIs to integrate universal search into their own Web pages or applications.

The Maintain Step

The maintain step ensures that search results are updated continuously. SES lets you gather from multiple Web sites and repositories, each on a different schedule. IMAP messaging servers, for example, can be updated continuously, while published content is gathered on a less frequent schedule. SES maintains content by providing easy, intuitive utilities that provide Administrators with an easy way to keep up with new content that is added through growth or acquisition.

Robust Connector Framework

Consumer search engines, like Google and Yahoo, index and search mainly HTML pages on web server. Enterprise Search Engines must also index Portals, Document Management Systems, custom applications and other software applications and systems. Oracle SES ships a family of built-in ‘connectors’ (Connectors are Java classes based on the SES plug-in API) for unlocking stored content in the most popular of these systems in use today.

The SES connector family provides access to documents that reside in the following proprietary systems and applications:

- Windows NT Filesystems (NTFS) -- NT fileshares can be indexed over a network connection and don’t have to be located on the SES host machine. SES provides strong access control by reading group and user access information and storing it in its search engine index.
- For SES installed on Unix operating systems, a small Agent process is installed in the same AD domain where the NT filesystem to be indexed is located. The agent sends content, metadata, and access control information to the connector in the SES machine (agent protocol is based on HTTP and can be encrypted via HTTPS). Microsoft IIS must be present for the agent to work.
- EMC Documentum Content Server – Indexes files in cabinets and folders of ContentServer DocBases. A native identity plug-in allows SES to show only those documents that a user has access

to according to permissions within Documentum. Efficient recrawls are supported – documents are only re-indexed if changed or moved within a Documentum.

- IBM Lotus Notes – Notes databases on IBM Lotus Notes Domino server instances (Notes Mail and custom applications planned for future release). The connector automatically navigates through all Notes databases on a Notes server instance. SES provides a Notes identity plug-in to use the Notes directory for authentication & validation of Notes-native users and groups.
- Microsoft Exchange – Indexes emails, attachments, calendar items and related metadata attributes in Exchange 2000 and 2003 stores. Efficient incremental recrawls are supported. Requires Microsoft IIS and 'Agent' software from Oracle (Agent, included with 10.1.8 release, sends content and metadata between Exchange host and SES host machines) from Oracle to be installed on the same Windows domain as the Exchange Server
- Microsoft Sharepoint

Oracle SES also searches across a number of Oracle-internal sources:

- OracleAS Portal page group, pages, and items
- Oracle Content Server (formerly Stellent, see section below for details)
- Oracle Collaboration Suite ContentServices and Calendar
- Oracle ContentDB – Folders, documents, and categories. Supports efficient re-crawls: Only documents with changed content, changed metadata/category metadata, and moved documents are re-indexed during incremental crawls.
- Oracle E-Business Suite 11i – Allows for crawling views, or queries, in Oracle database underlying 11i. Each record in the view or query is considered a separate document.
- Oracle Siebel 8.0 – RSS feeds.

All connectors are pre-configured and provide 'early binding' access control integration between SES and the legacy repository served by the connector (Early Binding means that the connector reads access control information for each document and provides this information to SES to store it in its search engine index). Many connectors are free of charge, but additional licensing is required for some major connectors.

Please see the document "Oracle Secure Enterprise Search 11g Connectors" on the [SES home page](#) Oracle Technology Network (OTN) for an up-to-date, complete list.

Security Plug-In Architecture

Secure Enterprise Search is directly integrated with third-party access control- and identity management solutions, including Microsoft's Active Directory. No synchronization of users or groups with Oracle Internet Directory is necessary. SES can directly access Active Directory (no extra coding required) through an authorization API and identity 'plug-in' architecture. SES ships plug-ins for Oracle's Internet Directory and Microsoft's Active Directory, among others. The architecture even

allows customers to build their own ‘identity plug-ins’ (supplies user and group information) for crawling sources with proprietary (non-LDAP) security schemes.

Performance and Scale

SES 11 is internally designed to scale. A single SES search server can serve up hundreds of gigabytes of content. SES servers can further be connected into federation clusters where they cooperate to serve up even more. To achieve performance at scale, Oracle introduced several innovations to the structure of its search engine index in order to reduce and eliminate unwanted I/O. We also enabled parallelism to SES query processing; it can now execute search requests simultaneously. This takes advantage of any parallel I/O capabilities in your hardware (for example, you might have multiple CPU cores, run SES on a server with multiple fast local drives, or have SAN storage with available parallel I/O bandwidth available).

Optimizations to the Structure of the Index

For very large document sets, search engine throughput is largely I/O bound. Search engines face a (famous within the industry) structural problem, called the ‘Long Tail’ law of search (Long Tail stands for the observation that the key words people search for in a given corpus do not occur with even frequency over time; instead a few terms occur over and over, followed by a ‘long tail’ of less and less popular terms. For example, the search keyword ‘Oracle’ or the word ‘Thanks’ occurs very frequently within the firm’s documents and emails, followed by other more unique terms like Stellent, BEA, Siebel etc. with lower and lower frequency). Caching does not provide much relief: The long tail distribution forces frequent cache hits for unusual (infrequently used) search terms. This problem is compounded by the regular content gathering cycles of the crawler, which invalidate any cached index parts regularly, typically overnight.

SES fetches and caches index blocks from disk in much larger, contiguous chunks and buffers than before, minimizing the number of times the engine has to go to disk when serving search requests. Oracle benchmark results show that fetching large parts of the index from disk in a single read operation (and caching them internally in equally large buffers) is several times more efficient than going to disk randomly to fetch small pieces of the index. Modern disk drives are extremely fast at reading large contiguous data sections, but take much longer to move their disk head to different section of the spindle to fetch smaller pieces here and there. Other changes in the layout of the index improve the performance of single word search queries.

SES automatically creates the newly optimized search engine structure during Release 11 installation. For customers upgrading from an earlier release (i.e. 10.1.8.4), Oracle offers an index migration tool, designed to be manually invoked after the upgrade procedure finishes.

SES Internally Parallelized To Leverage Multi-Disk, Multi-Core

Search queries can be executed in parallel to run faster on multi-core CPUs, on hardware with several fast disks drives, or if you have a high throughput network disk architectures (for example, SAN or NAS). The architecture works as follows:

- At install time, the search administrator provides a list of file paths that SES uses to partition its search engine index. Each path provided conceptually represents a separate disk drive or storage area. For best results, the number of partitions will be equal to the degree of parallelism in your disk I/O hardware. For ideal performance, the server machine hosting SES would have a 64bit capable, multi core CPU with multiple fast directly attached disk drives (for example, 4 core Opteron CPU with 12 x 3.5" 15K SAS drives) and enough RAM (Oracle recommends a minimum of 8GB, best would be 16 - 64 GB). But different storage technologies can be used, including network attached storage or an existing storage area network (SAN) – as long as your hardware is able to execute multiple disk requests in parallel fashion. An example would be SAN storage with Fiber Channel and sufficient available bandwidth.
- Each time SES and its embedded SES database start up, it creates and launches multiple processes ('slaves')
- Then, once SES receives search query requests from users, it splits the total work associated with serving each query into smaller jobs, giving each slave a small part of the total work for execution. A special algorithm ('partitioning engine') makes sure search queries are split up such that multiple disks, if present, are utilized. Multiple partitioning strategies are available. Initially, SES 11.1.2 will support hash based partitioning, there is a plan to add user attribute based strategies in a later release
- Upon receiving its query job, each slave will work on its assigned partition, then pass the results back to SES, which integrates all the results into a unified hit list

Concept Search and Result Clustering

Moving beyond keyword-based matching and singular hit result list presentation.

As the volume of information grows, even with high relevance the paradigm of keyword search starts reaching a plateau of diminishing returns. Users need advanced search techniques like the ability to look for concepts within their documents and cluster search results for iterative navigation.

SES includes the categorization and information-clustering (clustering is a technique for grouping objects based on similarity) technologies Oracle obtained from its earlier acquisition of Enterprise Search company TripleHop Technologies. What really makes this technology unique is what happens after you search. Instead of delivering thousands of search results in a long list, SES groups similar results together into clusters. Clusters help you see your search results by topic or by taxonomy category so you can zero in on exactly what you are looking for. Rather than scrolling through pages of search results, clusters help you find results you may have missed or that were buried deep inside the ranked result hit list.

Oracle's information clustering features:

- On-the-Fly and real-time topic and concept extraction from both crawled and federated sources, based on statistical analysis of the top 'N' documents (N is configurable) of the search results list. Oracle's algorithm is designed to strike a balance between the quality of topic clustering and the time

required to cluster. Exhaustively clustering all resulting hits for a given search request – millions of documents might be returned -- could take far longer than an end user might want to wait

- Clustering can be performed not only on the automatically extracted topics, but also on metadata items like author and creation date of a document. Search administrators can define their own cluster trees based on an agree-upon corporate taxonomy as metadata clusters can be hierarchical (e.g. Oracle -> Products -> Secure Enterprise Search)
- SES builds a topic hierarchy – a quick logical overview of the result set of a given search (see figure 6, below, for an example). Individual documents can be assigned to more than one cluster, and clusters can be on different topics. Cluster nodes with large document sets are further categorized into child cluster nodes, and a hierarchy is built to give the end user a quick logical overview of their search result hits
- The SES sample search application features an iterative navigation feature to dynamically expand topic clusters as search users navigate their way from a big picture view of all the content returned by their (often fuzzy) search request -- down to the specific piece of information pertinent to what they are actually looking for
- Oracle uses whole documents to form clusters rather than just title or description metadata, not just title and description metadata
- Topic clusters are enhanced with Natural Language processing. The words that appear in documents and in queries often have many morphological variants. Pairs of terms such as ‘computing’ and ‘computation’ will not be recognized as equivalent without some special processing. SES topic extraction utilizes so-called stemming algorithms, which reduce a word to its stem or root form (e.g. ‘compute’ and ‘computation’ are reduced to the single representative form ‘comput’). This means that different variants of a term can be conflated to a single representative form, reducing the number of distinct topics needed for representing a set of result hit documents. Different algorithms are used depending on the language, for example English and French use the well-known Porter algorithm.

Flexible parameters allow for customizing Oracle’s topic extraction algorithm:

- ‘Blacklist’/‘whitelist’: List of phrases/ words which should not be/ must be candidates for forming topic clusters if they appear among documents to be clustered. For example, a blacklist might contain entries like “site maps”, “term of use”, and “Oracle Corporation” (not a descriptive cluster name within Oracle)
- Minimum frequency counts and maximum number of one-word phrases, multi-word phrases, and sentences to be extracted
- Maximum number of cluster nodes at each level, levels of the cluster hierarchy and documents within one node

The Clustering / Topic Interface

The clustering capabilities can be embedded into end user applications from the Query Web Service API. The main interface to clustering is:

```
ResultContainer = doOracleOrganizedSearch (topN, duplicateControl,...)
```

It accepts the clustering request, along with several parameters and options. The output contains the cluster tree for the search request. Cluster trees can be returned in XML and JSON formats. An example of a cluster tree is shown below:

```
<cluster>
  <nodeset>
    <node id="1" name="all" level="1" size="100" leaf="0" keywords="all"/>
    <node id="1.4" name="java" level="2" size="99" leaf="0" keywords="java"/>
    <node id="1.4.1" name="data warehousing" level="3" size="38" leaf="0" keywords="technologies bi,
      data warehousing,linux .net office php security service"/>
    <node id="1.4.1.1" name="tutorials blogs" level="4" size="12" leaf="1" keywords="tutorials
      blogs">
      2773.,8031.,109.,8033.,806.,26940.,817.,8024.,8030.,2862.,8032.,8028.</node>
    <node id="1.4.1.2" name="stored procedure" level="4" size="4" leaf="1" keywords="stored
      procedure">
      4239.,4243.,2784.,4335.</node>
    <node id="1.4.1.3" name="miscellaneous" level="4" size="22" leaf="1">
      4017.,2836.,8029.,2767.,1502.,113814.,11731.,1138.,392.,2819.,2763.,1421.,221.,705.,
      7739.,2838.,2749.,2351.,2802.,1158.,15751.,15747.</node>
    </nodeset>
  </cluster>
```

Figure: Example cluster tree returned from API

The interface for clustering supports both a ‘rich client’ and a ‘thin client’ interaction mode. Rich clients make a single call to the SES server -- obtaining all of hit list, sorting, grouping, and clustering data associated with it – and are able to do sorting, grouping, and cluster navigation without any further round trips to the SES server. Thin clients rely on SES to manage pagination, sorting, grouping, and cluster navigation. For this mode, the clustering interface returns only a small chunk of the result hit list in a specific order.

Concept search and hit clustering supports the most common languages of Western- and Eastern European origin. Oracle plans support for Japanese, Chinese, and Korean in a subsequent release.

Powerful Search Query Syntax

Rich syntax exposes all the power of the Oracle Text platform, including Thesaurus expansion, fuzzy matching, and proximity search. Oracle provides rich syntax for performing query expansion, fuzzy search, Boolean, and grouping operations:

- Searches can be more like a programming language in supporting binary logical operators '&' (AND) and '|' (OR), and the parenthesis for grouping them together '(' '), so that you can do: '(Oracle & database) | (Enterprise & search)'
- Proximity search: "Oracle Database"~10 gives you matches with these two terms within 10 words of each other

Thesaurus & Alternate Query Terms

While false hits, or over-inclusiveness in full-text searching, is annoying, under-inclusiveness, or false misses, because of spelling variants, phrase variants, and the like is also a concern. Certain techniques can find word variants:

- Wildcard matchings are allowed: 'Ora*le Dat*base', 'Ora?le Dat?base'
- Fuzzy can sift through misspellings of a term: 'hallo~' will give you hits with 'hello'

Furthermore, customers can now define their own Thesaurus files and use them for Search. Thesauri take taxonomies and extend them to make them better by not only allowing subjects to be arranged in a hierarchy, but also allowing other relationships to be defined:

- Broader- (BT) and narrower term (NT): '<California' might find hits with 'San Francisco' and 'Los Angeles'. '>Ice cream' will give you 'desserts', 'unhealthy foods' and other related results
- Synonyms (SN) by preceding a word with '~'. For example '~car' might give you hits in cars, vehicles, automobiles, etc.

These capabilities are based on a Thesaurus that can be defined as an XML file and imported into Oracle Text (SES offers a command line tool for loading a Thesaurus into the Oracle Text engine of an SES server).

A related, but slightly different function is Alternate Keyword Expansion. Secure Enterprise Search has an alternate keyword feature that allows search administrators to suggest alternate search terms. For example, Oracle uses 'SES' and 'Secure Enterprise Search' interchangeably. To specify an alternate keyword, the SES administrator would enter both terms. The 10.1.8.2 release added an admin option to say 'auto expand'. When it is chosen, and when a user types in 'SES', hits with both 'SES' and 'Secure Enterprise Search' will be shown, with exact matches of 'SES' given higher relevancy.

This is different from Thesaurus based synonyms because an alternate keyword may not be a synonym and the query expansion is not initiated by the search user but by the search administrator.

However, at a certain point, extending the list of retrieved documents to encompass word variants will itself start resulting in false hits. An alternative to Boolean logic search terms is natural language "clustering" (see section above).

Attribute Shortcuts

Previously, the advanced search page was the only way to narrow searches by attributes like author or creation date. Now, that's a long detour simply adding one attribute constraint to your search. For example, you might searching for meeting notes written by your coworker Tom. Now, with the Attribute Shortcuts, you can simply say, 'meeting notes author:Tom' in the basic search box..

All the other operators above apply to the Attribute Shortcut:

- For synonym: 'safety rating title:~cars' gives you safety ratings for cars, vehicles, automobiles, etc
- For narrower/broader terms: 'weather report region:<California'
- Using Attribute Shortcut on numbers is very intuitive: 'digital cameras price:<500'.

Document Service Interface

The Document Service Interface turns SES into a powerful platform for building a customized search engine. It is a type of Java crawler plug-in which can be used to hook custom code into the crawler pipeline of Secure Enterprise Search. It is typically used to accept document from the SES crawler and perform custom operations such as:

- Add/change document attributes
- Change/filter the content of your documents
- Control whether or not each document should be indexed

The doc. service interface works with any supported data source type, and thus with all SES connectors, and has a wide range of potential applications, including clustering and classification. For example, your web assets might be manually tagged with metadata and you want your users to be able to restrict searches based upon the tags you have defined. The Document Service API can be used to filter your custom metatags from the document content. You can then pass the metadata to SES for categorizing your search results into your own Taxonomy.

Figure 6 below illustrates the flow of control. Several sequences of plug-in instances can form a pipeline. There is a global pipeline, but data source specific pipelines (one pipeline per source) can also be added.

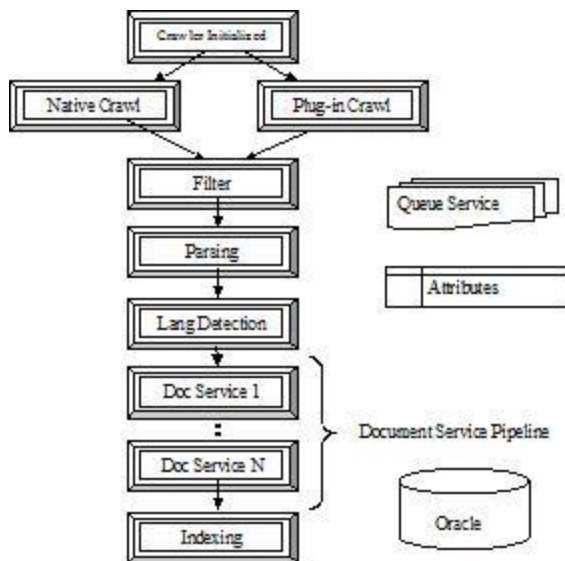


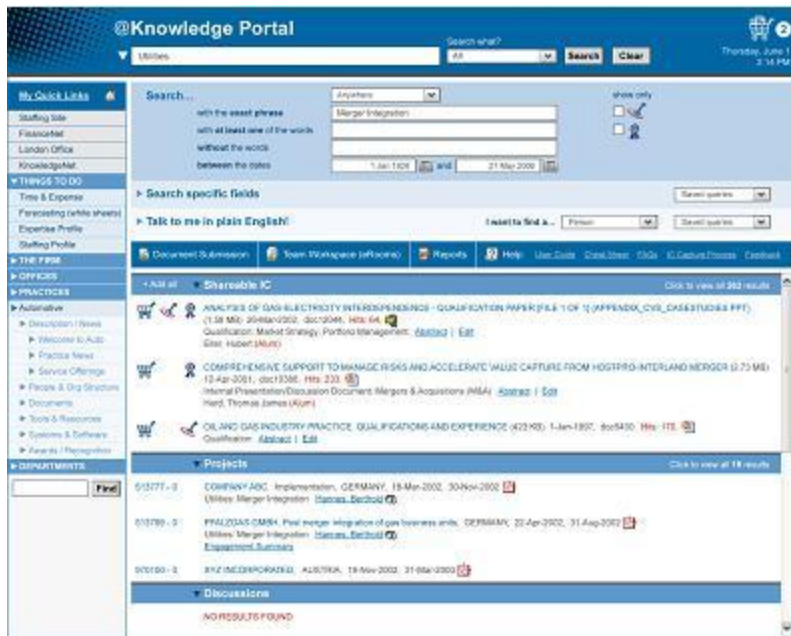
Figure illustrating of control flow in the document service API

Embedding SES as a Search Service

OEM partners and developers building and information applications or knowledge management software can embed SES as a 'search service' into their own software applications:

- Invoke searches from a search mask in your application via the SES Web service Query API
- Perform administrative actions, such as starting and stopping a crawler schedule or getting the index fragmentation level, remotely via Admin Web Service API
- Install SES silently with your software
- Extend SES metadata by pushing source-specific metadata to SES for searching
- Tune the relevancy of your search results based on application-specific characteristics. Use the Query Web Services API or a special parameter file, 'ranking.xml', to fine-tune the weights of default attributes (e.g. title, author) or add your own custom attributes and set weights for those attributes.

One example of a customer that embedded SES as central search service component of a knowledge management system (KMS) is AT Kearney, a professional services and management consulting firm headquartered in Chicago. Figure 11 below shows the search screen used by AT Kearney's worldwide consultants to find client deliverables (spreadsheets, client presentations) across multiple content sources. AT Kearney's KMS includes SES for search, a content management system for storing client deliverables, and screens that allow consultants to submit new client documents into the system.



Screenshot of AT Kearney's @Knowledge Portal, an example of a knowledge management system (KMS) that uses SES as embedded search service.

Other Features

The ‘Suggested Content’ feature lets you index and display real time content in the search results screen. A stylesheet can be applied to the content before it is displayed in the search result list.

SES allows you to run a silent installation (that is, an installation with preselected options and no interface). Silent installations make deployment on more than one computer much easier and can also be used for installation from a remote location (via command line). In a silent installation you supply Oracle’s Universal Installer with a response file and specify a ‘-silent’ flag on the command line. See the SES Administration guide for details.

You can supply an XML stylesheet to tailor the appearance of your search results in the SES search UI to a specific application or repository. Figure 12 shows an example of a search screen tailored to display custom metatags (‘session time’ and ‘event venue’) with each search result.

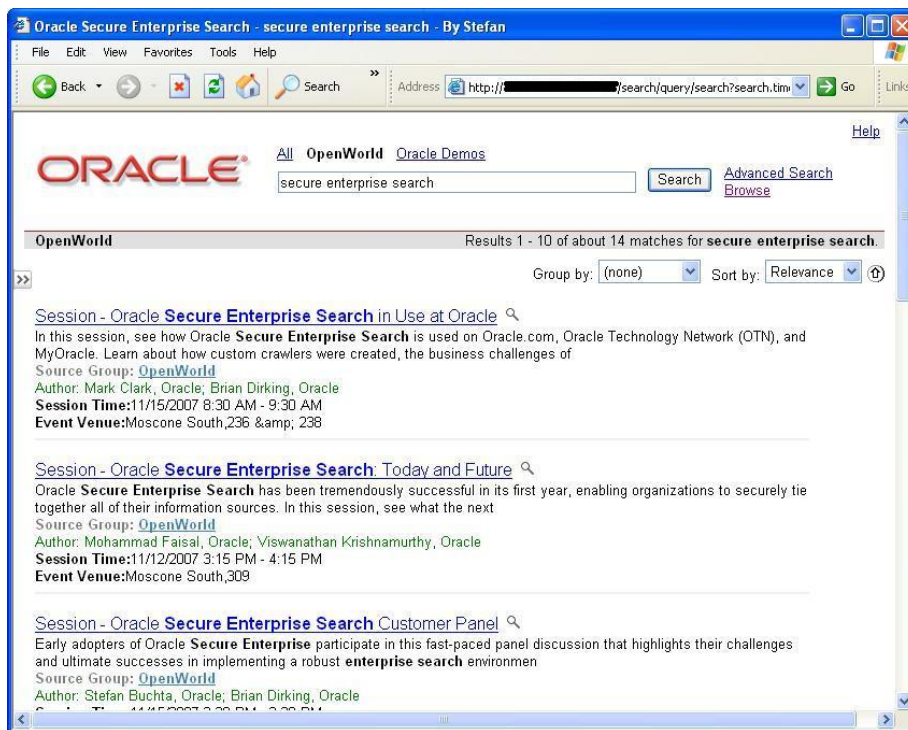


Figure illustrating hit list customization via XML stylesheet. Note how we display “Session Time” and “Event Venue” information directly below each search result.

Conclusion

The Enterprise Intranet is different from the Internet -- the information in it comes from many different types of sources; searches need to access password protected content; determining the importance of Intranet documents requires different techniques than on the Internet, and effective answers must often go beyond result hit lists. Secure Enterprise Search is built to bring to the Intranet

the information uplift users get on the Internet. By deploying Oracle's search solution, you can not only find information securely and effectively, mitigating information over-load, but also unlock the hidden intelligence that lies untapped in the deep Intranet.

Further Reading

- [1] SES Home page: <http://www.oracle.com/technetwork/search/oses/>
- [2] SES Administration Guide, ships with the product
- [3] White Paper "Enabling AutoVue Support in Oracle Secure Enterprise Search"
- [4] SES Datasheet
- [6] Whitepaper on SES Content Server Integration



White Paper Title
April 2013
Author: Stefan Buchta
Contributing: Mohammad Faisal

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com



Oracle is committed to developing practices and products that help protect the environment

Copyright © 2011, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. UNIX is a registered trademark licensed through X/Open Company, Ltd. 1010

Hardware and Software, Engineered to Work Together