Searching with Open Source products

Organizing your own world of information

- MnogoSearch
  – indexing and searching websites

- SphinxSearch
  – indexing and searching MySQL

FrOScon 2008
by Ralf Schwoebel, puzzler@tradebit.com

The meaning of search

Time is money:

Fast loading web pages convert better

Information is power:

Cross connecting information reveals potentially important data

Size matters:

Information stored online continues to explode: we are drawning in data

The real meaning of search

Split your mind:

Navigational
=> „I know what I want,
I forgot the location of document“

Research
=> „Show me what I should know“

Semantic search as the „new hype“ focusses on „ Disambiguation“ of terms.

Differences and meaning

The right tool for the task

mnoGoSearch:

(Group of) Websites (like Intranet)

Sphinx:

Huge databases and XML streams
=> specialized in full text search

Both tools need modifications, none is for semantic search.

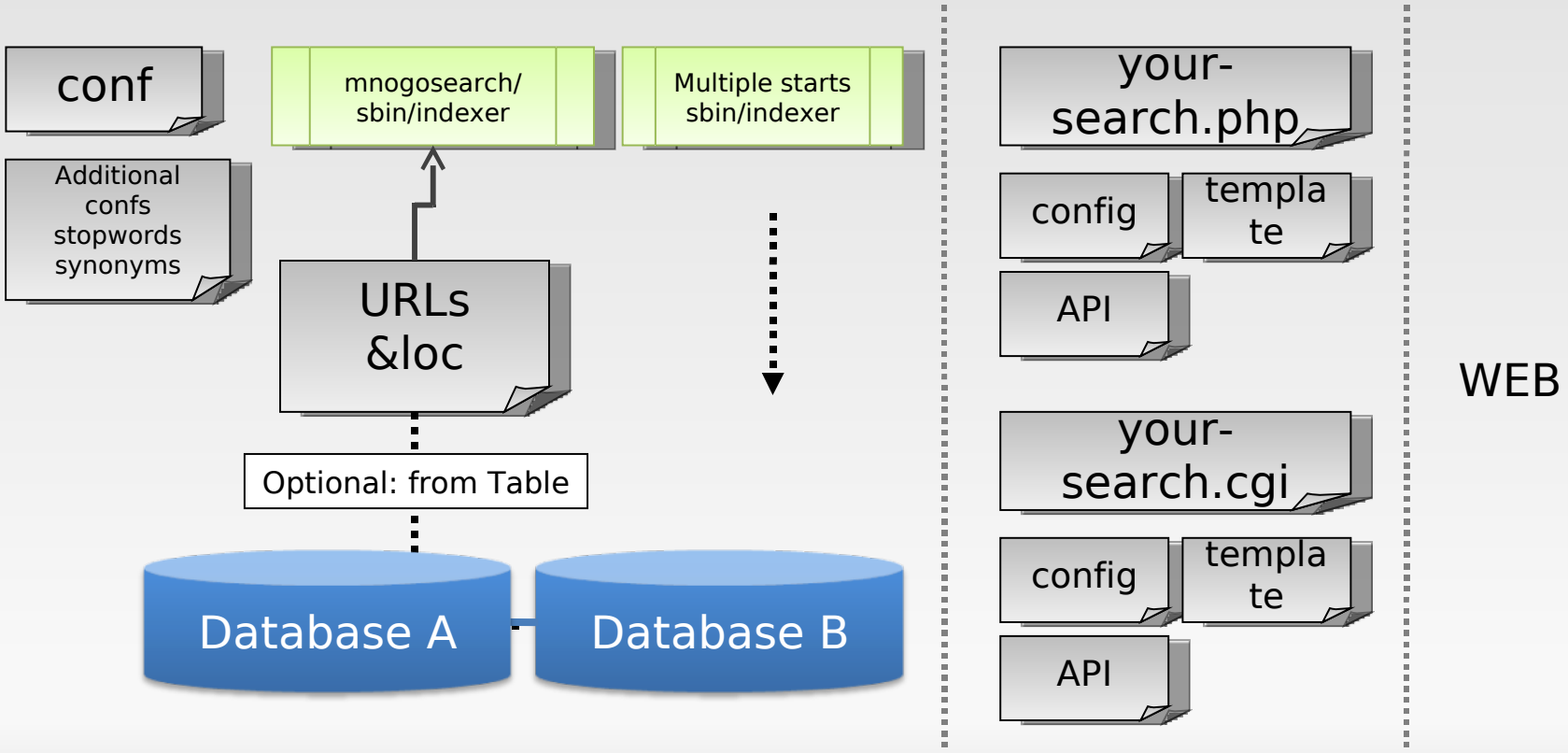
MnogoSearch

Feature Highlights

- Many backends: MySQL, Oracle, ODBC
- Spiders HTTP, HTTPS, NNTP, FTP
- and also FILES!
- Mirroring
- External parsers via modules (e.g. PDF)
- Fuzzy search: stemming, synonyms, etc.
- HTML templates for customization
- Various language APIs: PHP, CGI, C
- Authorization support (Intranet spider)

MnogoSearch

Architecture



MnogoSearch

Installation & configuration

„Linux-friendly“ TGZ download

./configure runs – tested on CentOS,
SuSE and ubuntu by myself
(autoconf, automake, gcc installed)

Installs to **/usr/local/mnogosearch/**
by default

MnogoSearch – meet the config

Shell config and PHP API have their own

```
# ls -lah /usr/local/mnogosearch/etc/indexer.conf
-rw-r--r-- 1 root root 39K Aug  6 12:05 /usr/local/mnogosearch/etc/indexer.conf
```

The purpose of the config entries are documented, e.g.:

```
# Currently supported DBType values are
# mysql, pgsql, mssql, oracle, ibase, db2, mimer, sqlite.
#
# MySQL users can specify path to Unix socket when connecting to localhost:

#mysql://mnogo:mnogo@localhost/mnogo/?socket=/var/lib/mysql/mysql.sock
```

MnogoSearch – meet the config

Important:

DBAddr <URL-style database description>
DBAddr mysql://mnogo:pwd1@d1/mnogo/?dbmode=blob&LiveUpdates=yes
DBAddr mysql://mnogo:pwd2@d2/mnogo/?dbmode=blob&LiveUpdates=yes
➔Cluster possible!

ServerTable <table_addr>
Load servers with all their parameters from the table specified in argument.
ServerTable mysql://mnogo:pwd@d1/mnogo/TBSERVER?srvinfo=TBSRVINFO
...

Section 3.
Mime types and external parsers.
➔ Plug in your PDF, PHOTO, Word parser!

...
Document sections.
➔ Set the weight of the sections like TITLE, HEADLINE or other tags

...
timeouts, identifications, proxies, etc.

MnogoSearch

The Power of REALMS

Realm regex (http://www\.)(.*)(\.yourname\.com/)(.*) file:/home/$2/htdocs/$4

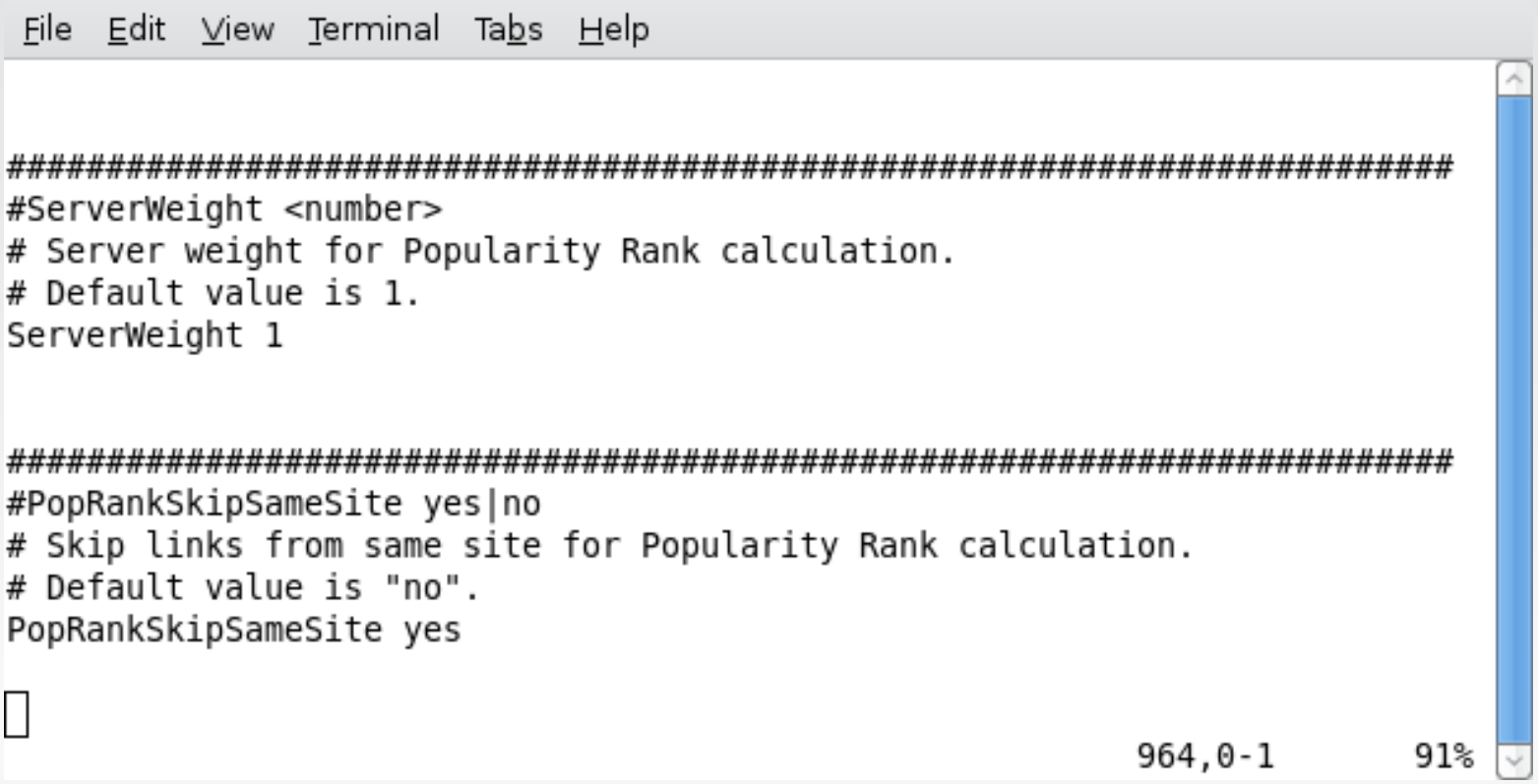
Mnogo spiders file systems as well!

Basically everything piped through EXEC:

Alias https:// exec:/usr/local/mnogosearch/etc/curl.sh?https://

MnogoSearch

Define weight and section to index



MnogoSearch

Go spider

help:	sbin/indexer --help
1st start:	sbin/indexer -Ecreate
spider:	sbin/indexer
rank & build search index:	sbin/indexer -Eblob
useful for tests:	sbin/indexer -c60

MnogoSearch

Show what you got!

d1:~ # /usr/local/mnogosearch/sbin/indexer -S

Database statistics [2008-08-23 18:51:20]

Status	Expired	Total	
0	2394	2394	Not indexed yet
200	17042	82503	OK
301	546	2192	Moved Permanently
302	613	4538	Moved Temporarily
303	3	104	See Other
304	973	3538	Not Modified
401	0	1	Unauthorized
403	17	85	Forbidden
404	534	1775	Not found
410	0	3	Gone
415	229	3013	Unsupported Media Type
500	1	8	Internal Server Error
503	135	4179	Service Unavailable
504	11	42	Gateway Timeout
Total	22498	104375	

MnogoSearch

Go find (php)

search.php:

…

> plug in your functions here

…

processes

$template_file='search.htm';

where you find the related config options

File Edit View Terminal Tabs Help

```
<!--
    This is default template file for mnoGoSearch 3.2
    (C) 1999-2002, mnoGoSearch developers team <devel@mnogosearch.org>
    Please rename to search.htm and edit as desired.
    See doc/README.templates for detailed information.
    You may want to keep the original file for future reference.
    WARNING: Use proper chmod to protect your passwords!
-->

<!--variables
# Database parameters are to be used with SQL backend
# and do not matter for built-in text files support
DBAddr  mysql://mnogo:mypwd@d1/mnogo/?dbmode=blob
DBAddr  mysql://mnogo:mypwd@d2/mnogo/?dbmode=blob

# Uncomment this line to enable search result cache
Cache no

# Uncomment this line if you want to detect and show clones
DetectClones yes

# Use proper local and browser charsets
# Examples:

LocalCharset    latin1
BrowserCharset  latin1

#LocalCharset   koi8-r
#BrowserCharset koi8-r

# For cache mode and built-in database
# you may choose alternative working directory
#VarDir /usr/local/mnogosearch/var

# Load stopwords file.  File name is either absolute
# or relative to /etc directory of mnoGoSearch installation.
#
StopwordFile stopwords/de.sl
StopwordFile stopwords/en.sl
#

#IspellUsePrefixes yes/no
:1                                                    1,1           Anfang
```

MnogoSearch

Go find (cgi)

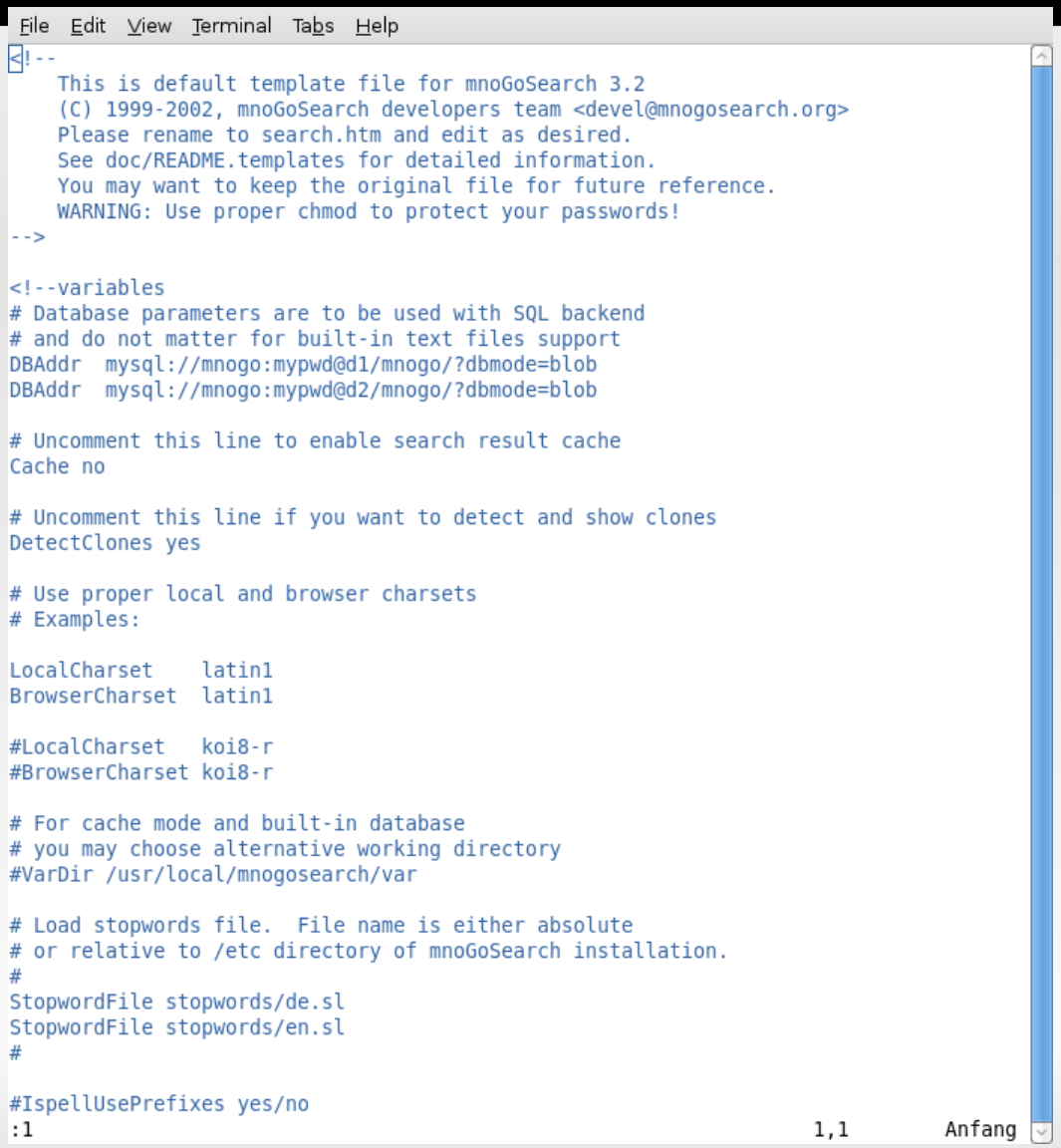
search.cgi:

…
 > binary – not easy to change
…

processes also

search.htm

where you find the related config options (again)



Conclusion

Mnogo Search is a fast and powerful search engine package and **especially handy for intranet searches** across different domains and servers. It is expandable and flexible.

How to handle and configure the package is **easy to learn**.

Professional support is available at reasonable prices directly from the coders.

SphinxSearch

Feature Highlights

- Speedy indexer (10MB/s of SQL/XML)
- Scalable: 100M docs / distributed search
- Native MySQL support (InnoDB, too!)
- Phrase proximity ranking
- Ranking weights
- Stopwords
- PHP API
- Compiles on most Linux flavours instantly
- Excellent x64 support

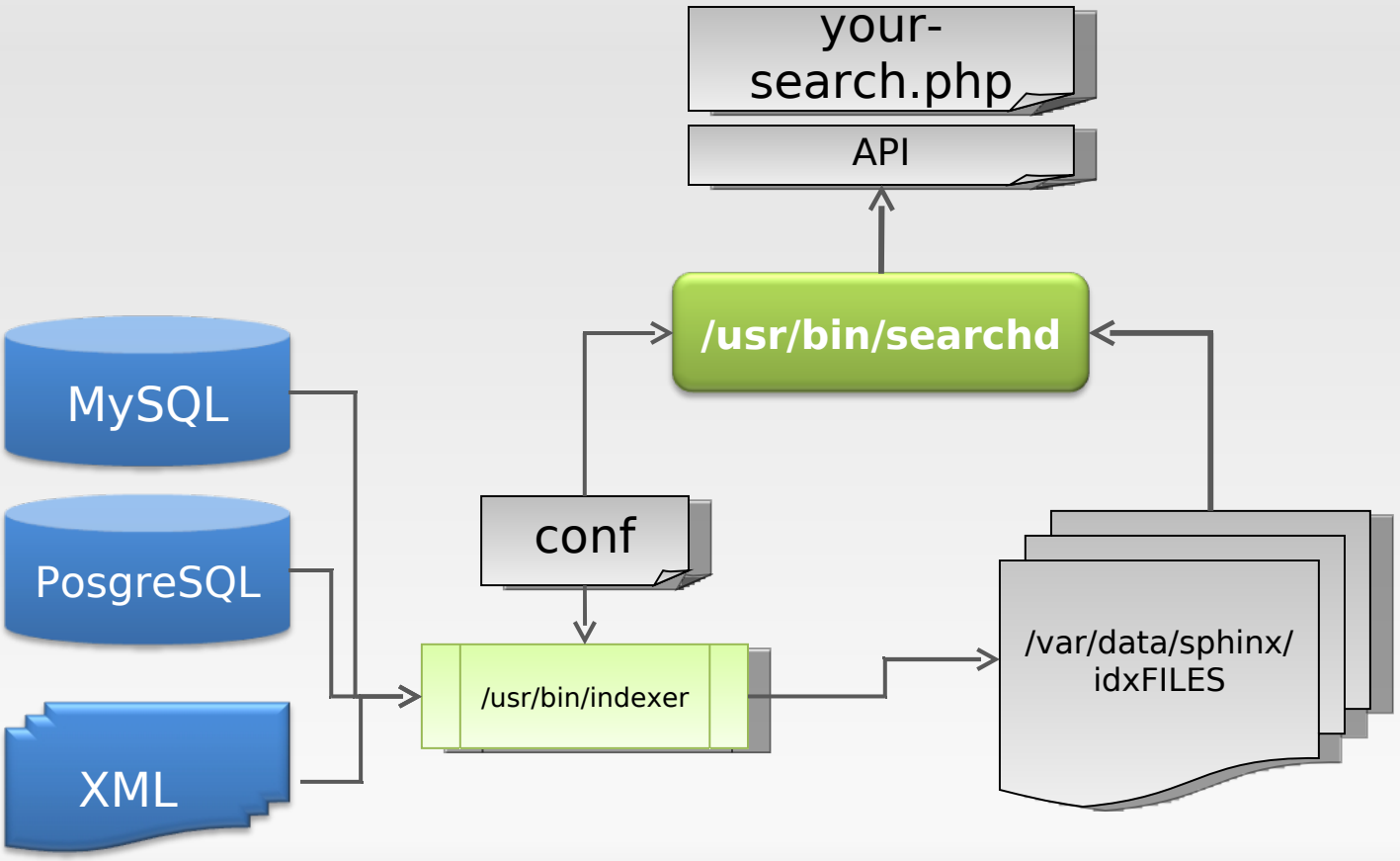
SphinxSearch

Why an extra MySQL fulltext search engine?

- WHERE like ‘%word%‘ statement not efficient
- MATCH AGAINST does not work in InnoDB mode
  but InnoDB needed for clustered DBs
- MATCH AGAINST gets REALLY slow on 10+ Million rows

SphinxSearch

Architecture



SphinxSearch

Installation and configuration

Compiles easily with „configure“ & „make“

Gives you serveral binaries:

- searchd
- indexer
- search (cli command to search your indices)

SphinxSearch – meet the config

Source definitons (example extract from sphinx.conf)

```
source srcTradebitEN
{
  type = mysql
  strip_html   = 1
  index_html_attrs =
  # some straightforward parameters for 'mysql' source type
  sql_host   = yourhost
  sql_user   = user
  sql_pass   = pw
  sql_db    = tradebit
  sql_port   = 3306 # optional, default is 3306
  sql_query_pre  =
  sql_query   = SELECT TB_INDEX, TB_NAME, TB_DESCR from TBFILES
  sql_query_range  = SELECT MIN(TB_INDEX),MAX(TB_INDEX) FROM TBFILES
  sql_range_step  = 1000
  sql_group_column = TB_FGRPID
  sql_group_column = TB_USERID
  sql_group_column       = TB_FLAGS
  sql_date_column  = date_added
  sql_query_post  =

  sql_query_post_index = REPLACE INTO sphix_counters_TradebitEN ( id, val ) VALUES ( 'max_indexed_id', $maxid )

  sql_query_info  = SELECT * FROM TBFILES,TBFILESTEXT_EN WHERE TBFILES.TB_INDEX=$id
                                AND TBFILES.TB_INDEX=TBFILESTEXT_EN.TB_INDEX
}
```

SphinxSearch

Index definitons (example extract from sphinx.conf)

```
# local index example
index idxTradebitEN
{
  source = srcTradebitEN
  path = /var/data/sphinx/idxTradebitEN

  # available values are "none", "inline" and "extern„
  docinfo = extern
  morphology = none

  stopwords =

  min_word_len = 1

  # known types are 'sbcs' (Single Byte CharSet) and 'utf-8'
  charset_type          = sbcs

  min_prefix_len = 0
  min_infix_len = 0
}
```

SphinxSearch

Go indexing

```
[root@sc ~]# indexer --config /etc/sphinx-localhost.conf --rotate idxTBUSERTAGSEN

Sphinx 0.9.8-release (r1371)
Copyright (c) 2001-2008, Andrew Aksyonoff

using config file '/etc/sphinx-localhost.conf'...

indexing index 'idxTBUSERTAGSEN'...

collected 1517615 docs, 22.9 MB
sorted 5.4 Mhits, 100.0% done

total 1517615 docs, 22875939 bytes
total 17.636 sec, 1297107.25 bytes/sec, 86051.52 docs/sec

rotating indices: succesfully sent SIGHUP to searchd (pid=26692).
```

SphinxSearch

Go find

```
[root@sc ~]# search --config /etc/sphinx-localhost.conf elvis
Sphinx 0.9.8-release (r1371)
Copyright (c) 2001-2008, Andrew Aksyonoff
using config file '/etc/sphinx-localhost.conf'...

index 'idxTradebitEN': query 'elvis ': returned 1000 matches of 2455 total in 0.000 sec

displaying matches:

1.              document=2418598, weight=3, tb_fgrpid=4008, tb_filetypeid=0,
                tb_userid=24435, tb_flags=1, date_added=Wed Nov 28 09:13:07 2007, tb_ispub=1
                TB_INDEX=2418598
....
16.             document=2840422, weight=3, tb_fgrpid=3014, tb_filetypeid=0,
                        tb_userid=28461, tb_flags=4, date_added=Tue Apr 22 00:00:00 2008, tb_ispub=1
                TB_INDEX=2840422
....
```

SphinxSearch

Integration with the PHP API

```
<?php
        require("include/sphinxapi.php");
        $sphinx_index = "idxTradebitEN“;
        $sphinx_filter = "TB_ISPUB";
        $sphinx_filter_args = array(1);
        $sphinx_limit = 10000;
        $strSphinxQuery = „your searchphrase“;
        $cl = new SphinxClient ();
        $cl->SetServer ( Config::$SRV_SPHINX, Config::$PORT_SPHINX );
        $cl->SetWeights ( array ( 1, 1, 1, 1 ) );
        $cl->SetFilter ( $sphinx_filter, $sphinx_filter_args );
        $cl->SetMatchMode ( SPH_MATCH_EXTENDED );
        $cl->SetSortMode ( SPH_SORT_RELEVANCE );
        $res = $cl->Query ($strSphinxQuery, $sphinx_index );
        if ($res) {
            if (count($res["matches"])>0) {
                $strSphingIdx = implode(",", array_keys($res["matches"]));
            }
        } else {
            print "Query failed: " . $cl->GetLastError() . "<BR>";
        }
        $filesincat = $res['total_found'];
?>
```

Does it scale? „Hell, yeah!“

Different concepts:

- Create „specialized“ indices on dedicated servers
- Partition/distribute the index on different drives:

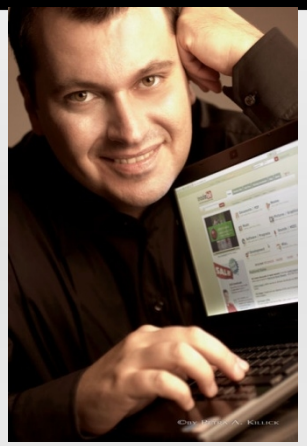
```
index dist1
{
    type = distributed
    local = chunk01
    agent = localhost:3312:chunk02
    agent = localhost:3312:chunk03
    agent = localhost:3312:chunk04
}
```

Conclusions

Sphinx is an excellent choice for full text searches on MySQL and XML sources.

Incremental indexing, fast response times and vast config options make the software „enterprise ready“ – backed by professional support.

Thank you



Thanks for watching,
contact me for questions

Ralf Schwoebel, CEO
Tradebit AG
puzzler@tradebit.com
http://www.tradebit.com/