Круто как в гугле. Поисковые сервисы на основе Sphinx

Владимир Федорков

DevConf, Москва 2014

Кто будет говорить?

- Работаю со сфинксом с 2006
- Performance geek
- Блог http://astellar.com
  - Не обновляется
- twitter @vfedorkov
  - Обновляется

О чем будем говорить

- Гугл приучил к хорошему
- Мало иметь информацию ее нужно уметь показывать
- Мало показывать, нужно уметь находить
- Мало уметь искать, нужно искать быстро
  - Угадывая чего хочет пользователь
- Посмотрим в чем Sphinx может нам помочь

Поисковые сервисы

- Автоподсказки
- Фесеты (Drill-down, narrowed search, faceted search)
- Коррекция ошибок ввода
- Управление релевантностью
- Связанные документы
- Гео-поиск
- Подсветка найденного

Почему нужен сфинкс?

- Базы данных умеют не все
- Сфинкс ищет быстрее БД
  - Раз так в 1000
    - Без учета шардинга
- Sphinx поддерживает SQL
- Быстрый, удобный, крутой

Чем сфинкс НЕ является

- НЕ плагин к MySQL
- НЕ требует MySQL
  - или другую базу для работы
- Умеет НЕ только SQL
- НЕ является заменой базы данных
  - НЕ во всех случаях
    - Пока еще?
    - Никогда!
      - См. OLAP vs OLTP vs Column vs FTS vs Webscale

Полнотекстовый поиск

- And, Or
  - hello | world, hello & world
- Not
  - hello -world
- Per-field search
  - @title hello @body world
- Field combination
  - @(title, body) hello world
- Search within first N
  - @body[50] hello
- Phrase search
  - “hello world”
- Per-field weights
- Proximity search
  - “hello world”~10
- Distance support
  - hello NEAR/10 world
- Quorum matching
  - "the world is a wonderful place"/3
- Exact form modifier
  - “raining =cats and =dogs”
- Strict order
- Document structure support
  - Sentence
  - Zone
  - Paragraph

Не полнотекстовый поиск

- В терминах SphinxQL
  - a = 5, a < 5, a > 5, a BETWEEN 3 AND 5
- В терминах SphinxAPI
  - SetFilter(), SetFilterRange(), SetFilterFloatRange()
- Условия применются *после* MATCH()

Не полнотекстовые атрибуты

- WHERE, ORDER, GROUP… не толко для интов
- Integers, floats, strings
  - 32 bit unsigned, 64 bit signed и bitfields
- MVAs
- JSON!
  - SELECT ALL(x>3 AND x<7 FOR x IN j.intarray)
  - SELECT j.users[3].address[2].streetname

Гео-поиск

- **GEODIST**(Lat, Long, Lat2, Long2)
- Поддержка для mi/km/m, deg/rad etc

SELECT *, **GEODIST(doc_lat, doc_long, $lat, $long)** as dist,

FROM sphinx_index ORDER BY dist DESC LIMIT 0, 20

Поиск по диапазону

- Цен
- Дат
- Рейтингов (и звездочек)
- INTERVAL(field, x0, x1, …, xN)

SELECT **INTERVAL(item_price, 0, 20, 50, 90)** as range, COUNT(*)
FROM my_sphinx_products GROUP BY range ORDER BY range ASC

Коррекция ошибок ввода

- “Did you mean” у Гугла
  - Коррекция слов
    - Britney vs Brittney vs Brittaney…
- Коррекция ввода
  - Забыл переключиться
- Должно быть основано на существующих в базе данных

Как сделать?

- Отдельный индекс + скрипт
- /misc/suggest/ в дистре сфинкса
  - suggest.conf
  - suggest.php
- Основан на поиске триграм
- Затем ранжирует матчи по Левенштейну
- Из коробки может уметь не все

Снипеты

- **CALL SNIPPETS** (or BuildExcerpts in the API)
- Конфиг
  - before_match “”
  - after_match “”
  - chunk_separator “…”
  - limit
  - around
  - force_all_words
  - …и т.п.

Релевантность

- Sphinx умеет ***expression based ranking***
  - OPTION ranker=expr('1000*sum(lcs)+bm25')
  - OPTION ranker=expr('700*sum(lcs)+bm25f(1.4, 0.8, {title=3, content=1}')
  - OPTION ranker=expr('$A*sum(lcs) + $B*atc + $C*bm25f + $D*…')

Дополнительные настройки

- Морфология
- blend_chars
- Перезапись запросов
  - Если поиск по фразам ничего не нашел:
    - включаем режим поиска слов
    - Ищем любые 2-3 слова
    - Ищем одно слово
    - Показываем рекламу

Похожие документы

- Используется основной индекс
- Используем кворум, сортируем по релевантности
  - “Sony NEX-5N”/0.3
- Включаем ручную сортировку
  - Та же категория => boost weight
  - Тот же ценник => boost weight
  - Тот же производитель => boost weight

Фасеты

- Не более чем группировка по признаку

```
mysql> SELECT id, WEIGHT(), ts, YEAR(ts), COUNT(*) as yr
    -> FROM lj1m WHERE MATCH('I love Sphinx')
    -> GROUP BY yr ORDER BY yr DESC LIMIT 5
    -> OPTION field_weights=(title=100, content=1);
+---------+----------+------------+------+----------+
| id      | weight() | ts         | yr   | count(*) |
+---------+----------+------------+------+----------+
| 7637682 |   101652 | 1112905663 | 2005 |       14 |
| 6598265 |   101612 | 1102858275 | 2004 |       27 |
| 7139960 |     1642 | 1070220903 | 2003 |        8 |
| 5340114 |     1612 | 1020213442 | 2002 |        1 |
| 5744405 |     1588 |  995415111 | 2001 |        1 |
+---------+----------+------------+------+----------+
```

Особеннсти SQL

- Расширение MySQL протокола
  - WITHIN GROUP ORDER BY
  - GROUP <N> BY etc
- Поисковые расширения
  - OPTION
  - SHOW META
  - CALL SNIPPETS
  - CALL KEYWORDS

В чем разница

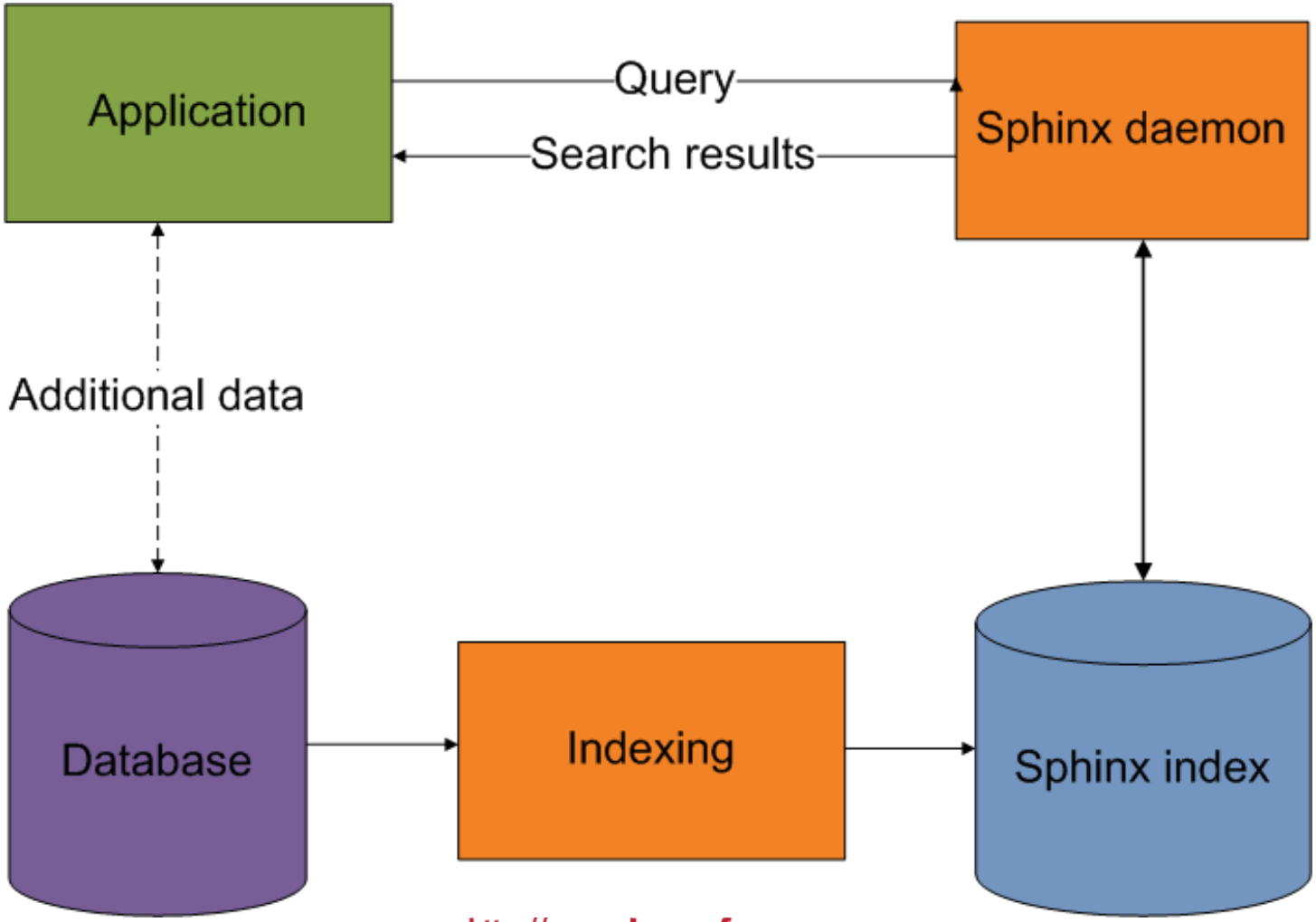
```
mysql> SELECT id, ...
    -> FROM myisam_table
    -> WHERE MATCH(title, content_ft)
    -> AGAINST ('I love sphinx') LIMIT 10;
...
10 rows in set (1.18 sec)
```

MySQL

```
mysql> SELECT * FROM sphinx_index
        -> WHERE MATCH('I love Sphinx') LIMIT 10;
...
10 rows in set (0.05 sec)
```

Sphinx

Как работает?

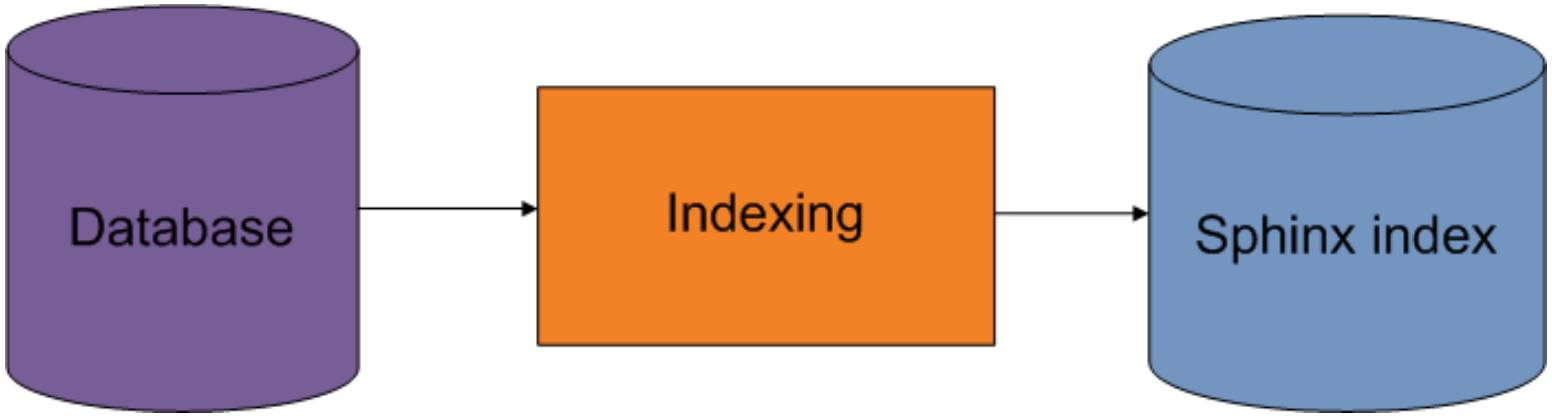


Как настроить?

- Запустить сфинкс
- Подключиться
- Запилить запрос

Конфигурация

- Где брать данные
- Как индексировать
- Где и как хранить



Откуда брать

- MySQL
- PostgreSQL
- MSSQL
- Oracle
- ODBC source
- XML pipe
- CSV pipe
- …

Источник

```
source data_source
{
        …
        sql_query = \
                SELECT id, channel_id, ts, title, content \
                FROM mytable

        sql_attr_uint           = channel_id
        sql_attr_timestamp      = ts
        …
}
```

...или так

```
source data_source
{
  type        = mysql
  sql_host    = localhost
  sql_user    = myuser
  sql_pass    = mybiggestsecret
  sql_db      = test

  sql_query_pre = SET NAMES utf8
  sql_query     = SELECT id, channel_id, ts, title, content \
                  FROM mytable WHERE id>=$start and id<=$end
  sql_attr_uint = channel_id
  sql_attr_uint = ts

  sql_query_range  = SELECT MIN(id),  MAX(id) FROM mytable
  sql_range_step   = 1000
}
```

Индекс

```
index my_sphinx_index
{
        source                  = data_source
        path                    = /my/index/path/idx

        html_strip              = 1
        morphology              = lemmatize_en_all
        stopwords               = stopwords.txt
}
```

Индексатор

```
indexer
{
    # upto 2047M
    mem_limit       = 512M

    # for a busy system
    max_iops  = 40
    max_iosize = 1048576
}
```

Демон

```
searchd
{
        listen = localhost:9312
        listen = localhost:9306:mysql4

        query_log               = query.log
        query_log_format        = sphinxql

        pid_file                        = searchd.pid
}
```

Сфинкс

```
$ mysql -h 127.0.0.1 -P 9306
Welcome to the MySQL monitor.  Commands end with ;
or \g.
Your MySQL connection id is 1
Server version: 2.1.0-id64-dev (r3028)

Type 'help;' or '\h' for help. Type '\c' to clear
the current input statement.

mysql>
```

Важно знать!

- Мастер класс по сфинксу будет завтра
- Шодан будет сегодня
  - Рассказывать про внутренности NoSQL
- Главное не забыть про обед
  - С 15:30 до 16:00
    - Жрать придется быстро!

ВОПРОСЫ!

Если что-то осталось не понятным

- Поймать меня сегодня на конференции
- Написать мне через мой сайт
  - http://astellar.com/contact-me/
- Задать вопрос в группе в контакте
  - http://vk.com/sphinxsearch

СПАСИБО!