Движки полнотекстового поиска в Firebird

Сергей Волков
Системный архитектор

FIREBIRD CONF 2024 | РЕДСОФТ

Полнотекстовый поиск (FTS)

- поиск по содержимому документов или для СУБД по полям;
- поиск по термам, а не по однозначному соответствию;
- операторы модификации: увеличения веса термы, обязательность или отсутствие термы и т.д.;

Полнотекстовый поиск (FTS)

- Apache Lucene (java) интеграция от Ред Софт;
- CLucene (C++) интеграция от IBSurgeon;
- Sphinx (C++) не поддерживается;
- *Другие*

CLucene — IBSurgeon FTS UDR

- Разработана Денисом Симоновым, разработку финансировал IBSurgeon.
- Windows — бинарники.
- Linux — исходники.

Почему под Linux исходники?

ОС	liblucene++	libboost
RedOS 7.3	3.0.7-17	1.73
RedOS 8.0	3.0.7-33	1.78
Astra Linux 1.7	3.0.7-10	1.67
Alt Linux 10	3.0.8	1.76
Ubuntu 24.04	3.0.9	1.83
Debian 12	3.0.8-6	1.74

Clucene — настройка

- Установка.
- Настройка в fts.ini указать для баз директорию хранения полнотекстового индекса
- fts.ini находиться в директории ${FIREBIRD_HOME}

Clucene — создания индекса

- Пакет FTS$MANAGMENT.
- FTS$CREATE_INDEX('ISSUE', 'ISSUE', 'RUSSIAN', 'RDB$DB_KEY'); -- Создание индекса
- FTS$ADD_INDEX('ISSUE', 'SUBJECT'); -- Добавляем поле к индексу
- FTS$REBUILD('ISSUE'); --Индексация

Clucene — поиск

- FTS$SEARCH('ISSUES', 'Тестирование', 10000, false);
- Возвращает параметры:
- FTS$RELATION_NAME — имя таблицы;
- FTS$KEY_FIELD_NAME — имя ключевого поля;
- FTSDB_KEY, FTSID, FTS$UUID — значение ключевого поля;
- FTS$SCORE — степень соответствия поисковому запросу;
- FTS$EXPLANATION — объясняет результаты поиска.

Clucene — актуализация данных

- Пакет FTS$TRIGGER_HELPER
- FTS$MAKE_TRIGGERS
- FTS$MAKE_TRIGGERS_BY_INDEX
- FTS$MAKE_ALL_TRIGGERS

FBJava Lucene library

- Разработана отделом разработки СУБД, компании РЕД СОФТ
- Windows, Linux — бинарники (jar).
- Поставляется в промышленной редакции СУБД Ред База Данных, но может работать в любой редакции с FBJava.

FBJava Lucene library — настройка FBJava

- В plugins.conf раскомментировать секции Plugin=Java и Config=Java_config

```
Plugin = JAVA {
        Module = $(dir_plugins)/fbjava
        Config = JAVA_config
}
```

```
Config = JAVA_config {
        JavaHome = /usr/lib/jvm/java-11-openjdk-amd64
        SecurityDatabase = $(this)/java-security.fdb
        JvmArgsFile = $(this)/jvm.args
        JarDirs = $(this)/jar
}
```

- В fbjava.yaml указать путь к Java-библиотекам FTS:
- classpath:

    - $(root)/jar/fts/*.jar

FBJava Lucene library — настройка java-security

- Выполнить скрипт misc/fts_permissions.sql на БД java-security.fdb:
- ./bin/isql -u SYSDBA -p *** -i misc/fts_permissions.sql localhost:/opt/RedDatabase/java-security.fdb

FBJava Lucene library — настройка jvm.args

- -Dfts.directory=/data/fts — директория с файлами fts;
- -Dfts.disableGUID=true — отключает создание под директории с GUID БД;
- -Xmx16G — ограничивает размер heap 16GB.

FBJava Lucene library — создания индекса

- FTS$CREATE_INDEX('ISSUES', 'RUSSIAN', '', false) — создаем индекс.
- FTS$ADD_FIELD_TO_INDEX('ISSUES', 'ISSUES', 'SUBJECT');
- FTS$ADD_FIELD_TO_INDEX('ISSUES', 'ISSUES', 'DESCRIPTION');
- FTS$ADD_FIELD_TO_INDEX('ISSUES', 'JOURNALS', 'NOTES') - добавляем поля в индекс.
- FTS$FULL_REINDEX('ISSUES') — выполнить полную переиндексацию.

FBJava Lucene library — поиск

- FTS$SEARCH('ISSUES', null, 'Тест обновления', 100)
- Возвращает параметры:
- FTS$ROW_ID — RDB$DB_KEY,
- FTS$SCORE — оценка соответствия,
- FTS$RELATION — имя таблицы в котором найдено запись,
- FTS$HIGHLIGHT — фрагмент текста с подсвеченным результатом поиска (теги <B>, </B>).

FBJava Lucene library — актуальность FTS-индекса

- FTS$TRIG_X — триггер вызывается после вставки, пишет в таблицу FTS$POOL.
- FTS$POOL:
- FTS$ROW_ID — RDB$DB_KEY изменившейся записи
- FTS$STATUS — статус изменившейся записи
- Индексирует FTS$SEARCH или запущенный демон

Сравнение производительности - объект

- БД redmine 3,5ГиБ;
- Таблица issues (обращения) — 171 тыс. записей, средний размер записи 197 байт;
- Таблица journals (комментарии) — 1,5 млн. записей, средний размер записи 51 байт;

Сравнение производительности - цифры

	IBSurgeon FTS UDR	FBJava Lucene library	containing	like
Время индексации	01:39	11:36		
Использование ОЗУ при индексации, МБ	270	7 263		
Размер Индекса	155	398		
Время поиска по фразе «разыскное дело», 100 первых результатов, сек	0,016	0,061	3,346	0,307
Время поиска по фразе «разыскное дело», все записи, сек	0,311	4,65	8,487	17,635
Количество записей возвращаемых при поиске по фразе «разыскное дело»	13 218	14 147	307	7 924
Использование ОЗУ при поиске, МБ	70	737		

Фишки Ред Базы Данных

- Функции работы с файлами
- Работа с pdf, odt, docx и т.д.
- OCR

Проблемы

- Репликация RDB$DB_KEY не может быть ключом.
- После backup/restore индекс нужно перестравать
- Java — высокое потребление ресурсов, C++ — ограниченный функционал

Спасибо за внимание!