

Використання Apache Solr для реалізації повнотекстового пошуку у комерційних системах

Виконав
Студент КН-1 МП
Кривонос А.А.

Керівник:
Яремко С.А.



Актуальність і мета курсової роботи

З розвитком технологій виникає необхідність у навігації по величезних обсягах даних. Звичайний інформаційний пошук засобами СКБД не завжди задовольняє всі потреби, зокрема у швидкості та гнучкості. Саме тому існує ряд спеціалізованих рішень, таких як Apache Solr, ElasticSearch, Lunrjs, Apache Nutch тощо.

Метою курсової роботи є дослідження теми повнотекстового пошуку і платформи Apache Solr, а також її застосування для реалізації пошуку у комерційних системах, зокрема у музичних стримінгових сервісах.



Коротко про розділи курсової роботи

- ❑ Розділ 1. Аналіз предметної області
 - ❑ Визначення повнотекстового пошуку
 - ❑ Бібліотека Apache Lucene
 - ❑ Загальні відомості про Apache Solr
 - ❑ Опис альтернативної платформи Elasticsearch та її порівняння з Apache Solr
- ❑ Розділ 2. Огляд інструментів розробки
 - ❑ Детальний розбір понять Apache Solr
 - ❑ Опис інших технологій (мови Python та використаних бібліотек)
- ❑ Розділ 3. Реалізація системи
 - ❑ Опис вхідних даних та схеми
 - ❑ Функціональні особливості системи
 - ❑ Архітектура
 - ❑ Тестування та опис можливостей для покращення



Apache Solr

Apache Solr - це open-source пошукова платформа. Її розробка розпочалась у 2004 році. Заснована на бібліотеці Apache Lucene.

Серед основних можливостей Solr є повнотекстовий та фасетний пошук, індексування у реальному часі, підтримка кластеризації та масштабування, та зручна інтеграція з базами даних.

Одним із найбільших конкурентів є Elasticsearch, який також заснований на Apache Lucene.



Інструменти розробки

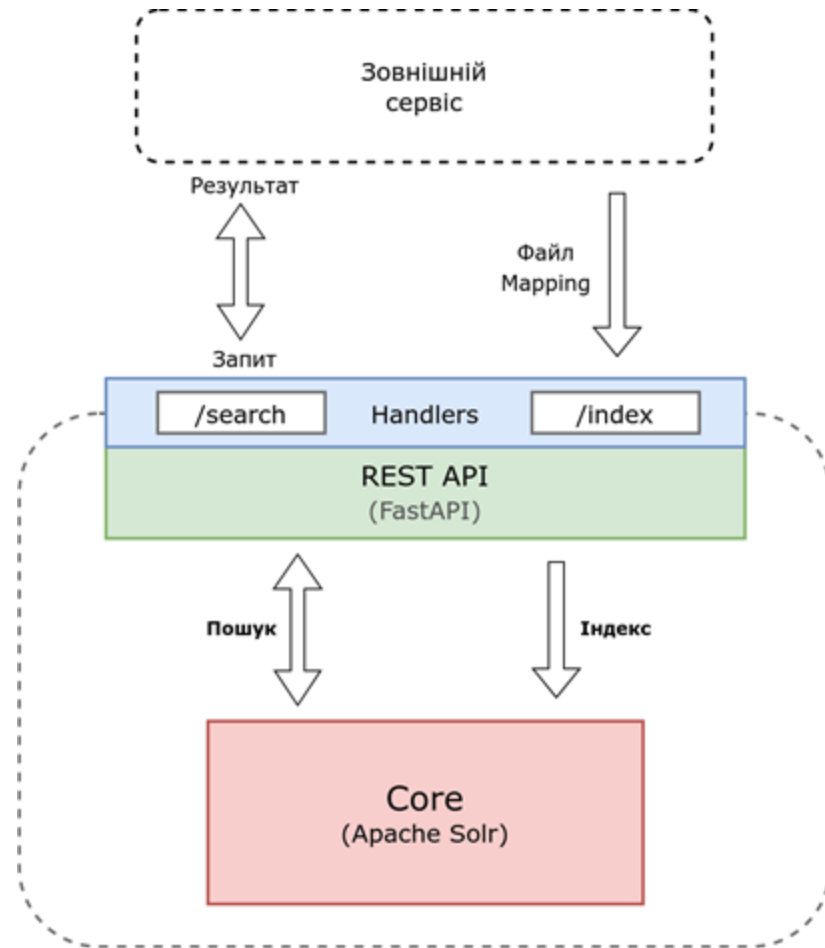
- Apache Solr
- Python
 - FastAPI
 - pysolr
 - requests

Архітектура системи

Ядром системи виступає Apache Solr

Для взаємодії зовнішніх сервісів імплементовано REST API, який має два основних ендпоінти:

- /search - для пошуку
- /index - для оновлення даних





Основні функціональні особливості

- Фільтрація зайвих даних, які подаються на індексацію
- Вилучення дублікатів
- Один єдиний ендпоінт для пошукових запитів по полям:
 - Назва пісні
 - Виконавець
 - Текст пісні
- Нечіткий пошук (часткове нівелювання допущення помилок у запиті)
- Маппінг полів (можливість інтеграції з іншими сховищами даних, де поля мають інші назви)



Тестування

Тестування було проведено для різних юзкейсів:

- Коли користувач знає точну назву виконавця та пісні
- Коли користувач знає назву виконавця та декілька слів з пісні
- Коли користувач знає приблизні слова з пісні
- Коли користувач вводить запит з помилками

Інструмент для тестування: **Postman**

Тестування

Точна назва пісні і виконавця:

“Three days grace chalk outline”

(Three Days Grace – Chalk Outline)

GET localhost:8000/search?query=three days grace chalk outline

Params Authorization Headers (6) Body Pre-request Script Tests Settings Cookies Code

Query Params

KEY	VALUE	DESCRIPTION	...	Bulk Edit
<input checked="" type="checkbox"/> query	three days grace chalk outline			
Key	Value	Description		

Body Cookies Headers (4) Test Results Status: 200 OK Time: 55 ms Size: 5.26 KB Save Response

Pretty Raw Preview Visualize JSON

```
1 {
2   "doc_found": 10353,
3   "start": 0,
4   "end": 10,
5   "results": [
6     {
7       "artist_name": "three days grace",
8       "track_name": "chalk outline",
9       "release_date": 2012,
10      "genre": [
11        "pop"
12      ],
13      "lyrics": "curse cross beat ones open shatter ones think love leave like chalk outline sidewalk wait
14        rain away away come scene crime dead speak leave leave lyric commercial",
15      "id": "41b408ae-bfal-4676-af53-c9da3ad0f04a",
16      "_version_": 1735172201975382018
17    },
18    {
19      "artist_name": "three days grace",
20      "track_name": "over and over",
21      "release_date": 2006,
22      "genre": [
23        "pop"
24      ],
25      "lyrics": "feel bring blame try away chase fall feel like stay dran null away chase fall fall thoughts
```

Тестування

Назва виконавця і кілька слів з пісні:

“lorde gold teeth grey trippin”

(Lorde – Royals)

The screenshot shows a REST client interface with a GET request to `localhost:8000/search?query=lorde gold teeth grey`. The response is a JSON object with the following structure:

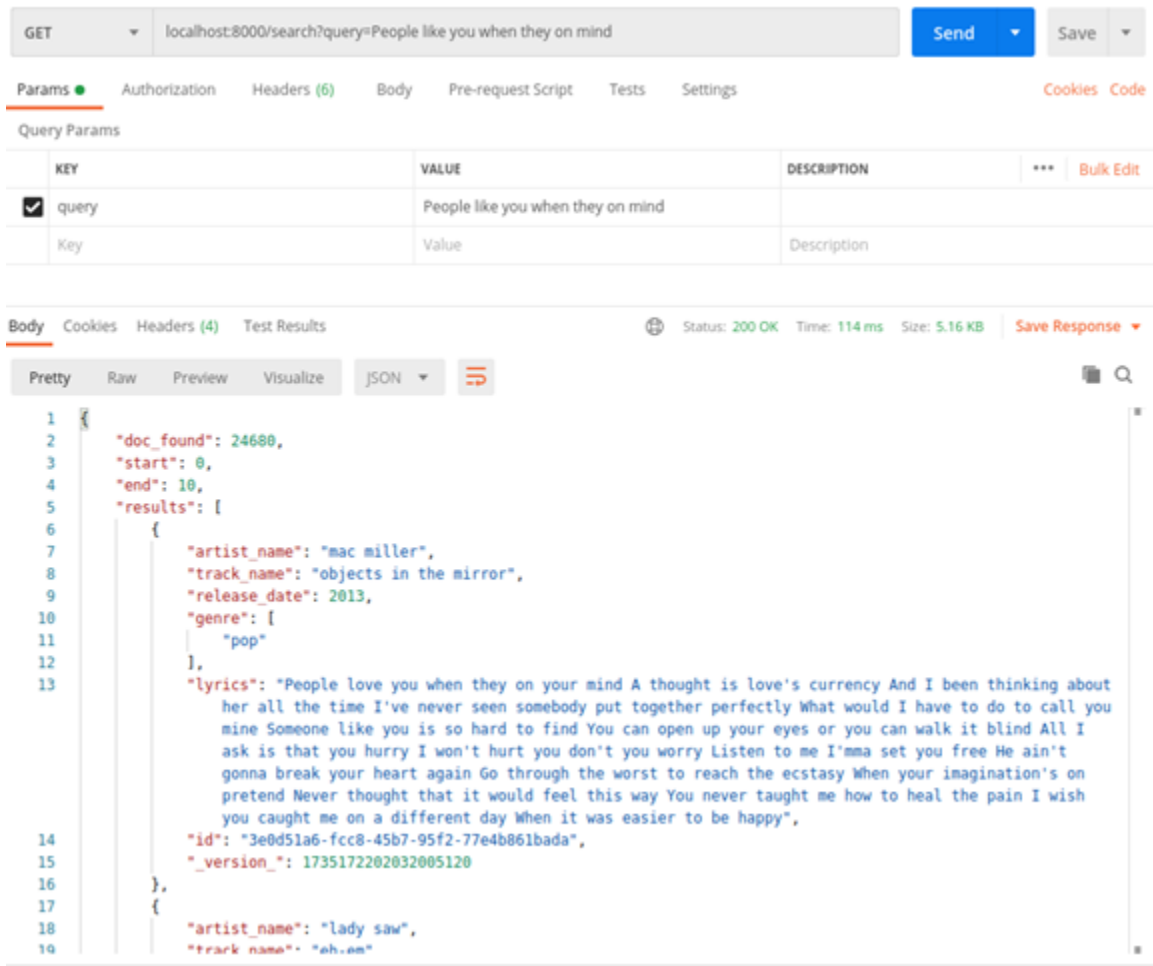
```
1  {
2    "doc_found": 14600,
3    "start": 0,
4    "end": 10,
5    "results": [
6      {
7        "artist_name": "lorde",
8        "track_name": "royals",
9        "release_date": 2013,
10       "genre": [
11         "pop"
12       ],
13       "lyrics": "see flesh teeth wed ring movies proud address tear postcode envy song like gold teeth grey
14         trippin bathroom bloodstains ball gown trashin hotel room care drive cadillacs dream everybody like
15         maybach diamonds timepiece plan islands tiger gold leash care aren catch affair royals royals blood
16         kind luxe crave different kind buzz ruler ruler queen baby rule rule rule live fantasy friends
17         crack code count dollars train party know know fine come money song like gold teeth grey trippin
18         bathroom bloodstains ball gown trashin hotel room care drive cadillacs dream everybody like maybach
19         diamonds timepiece plan islands tiger gold leash care aren catch affair royals royals blood kind
20         luxe crave different kind buzz ruler ruler queen baby rule rule rule live fantasy ohoh bigger
21         dream queen ohoh life great care aren catch affair royals royals blood kind luxe crave different
22         kind buzz ruler ruler queen baby rule rule rule live fantasy",
23       "id": "3b2dadcb-1f0c-44ee-9c13-cd463b4638c6",
24       "_version_": 1735172202025713666
25     }
26   ]
27 }
```

Тестування

Приблизні слова з пісні:

“People like you when they
on the mind”

(Mac Miller – Objects in the Mirror)



GET localhost:8000/search?query=People like you when they on mind

Params Authorization Headers (6) Body Pre-request Script Tests Settings Cookies Code

Query Params

KEY	VALUE	DESCRIPTION	...	Bulk Edit
<input checked="" type="checkbox"/> query	People like you when they on mind			
Key	Value	Description		

Body Cookies Headers (4) Test Results Status: 200 OK Time: 114 ms Size: 5.16 KB Save Response

Pretty Raw Preview Visualize JSON

```
1
2  "doc_found": 24680,
3  "start": 0,
4  "end": 10,
5  "results": [
6    {
7      "artist_name": "mac miller",
8      "track_name": "objects in the mirror",
9      "release_date": 2013,
10     "genre": [
11       "pop"
12     ],
13     "lyrics": "People love you when they on your mind A thought is love's currency And I been thinking about
her all the time I've never seen somebody put together perfectly What would I have to do to call you
mine Someone like you is so hard to find You can open up your eyes or you can walk it blind All I
ask is that you hurry I won't hurt you don't you worry Listen to me I'mma set you free He ain't
gonna break your heart again Go through the worst to reach the ecstasy When your imagination's on
pretend Never thought that it would feel this way You never taught me how to heal the pain I wish
you caught me on a different day When it was easier to be happy",
14     "id": "3e0d51a6-fcc8-45b7-95f2-77e4b061bada",
15     "_version_": 1735172202032005120
16   },
17   {
18     "artist_name": "lady saw",
19     "track_name": "ah.ah"
```

Тестування

Запит з помилками:

“tom odell anozer lov“

(Tom Odell – Another Love)

GET localhost:8000/search?query=tom odell anozer lov

Params Authorization Headers (6) Body Pre-request Script Tests Settings Cookies Code

Query Params

KEY	VALUE	DESCRIPTION	...	Bulk Edit
<input checked="" type="checkbox"/> query	tom odell anozer lov			
Key	Value	Description		

Body Cookies Headers (4) Test Results Status: 200 OK Time: 65 ms Size: 5.45 KB Save Response

Pretty Raw Preview Visualize JSON

```
parade order order obey duty order hurry home come station jump train march double lover roses
entwine arm arm arm surrender girl love soldier cause away soldier duty answer call loudest come
station jump train march double lover roses entwine arm arm arm surrender",
  "id": "bc678dd6-a921-4cf8-bb90-caalefd7e8d1",
  "_version_": 1735172200353234944
},
{
  "artist_name": "tom odell",
  "track_name": "another love",
  "release_date": 2013,
  "genre": [
    "pop"
  ],
  "lyrics": "wanna know care cold know bring daffodils pretty string like spring wanna kiss feel right
tire share nights wanna wanna tear tear tear tear somebody hurt wanna fight hand break time voice
fuck rude word know lose sing song sing heart wanna wanna learn tear tear tear tear need heart think
wanna sing song sing heart wanna wanna fall tear tear tear tear",
  "id": "6825330d-cbc8-46ae-a16a-2406281ecda4",
  "_version_": 1735172202030956544
},
{
  "artist_name": "j-kwon",
  "track_name": "they ask me",
  "release_date": 2004,
```



Можливості для покращення системи

- Додати синхронізацію з зовнішніми БД
- Зробити інтеграцію з рекомендаційними системами (додавати динамічні ваги для улюблених жанрів та виконавців)
- Додавати фасетинг для підбивання статистики по окремим жанрам
- Реалізувати підтримку додаткових полів (наприклад, параметрів звуку)



Висновки

- Було досліджено тему повнотекстового пошуку та платформу Apache Solr
- Реалізовано систему, яка виконує пошук по музичній колекції
- Перевагами Apache Solr є швидкість пошуку, велика кількість функціональних можливостей, підтримка кластеризації та можливість динамічної конфігурації через REST API
- Недоліки Apache Solr: складність відлагодження при виникненні помилок, плутанина в документації через різні версії.
- Apache Solr займає свою нішу, але Elasticsearch дедалі більше випереджає її за багатьма показниками