Система конвеерного индексирования данных в Apache Solr

Б17-191-1 Верещак В.Д.

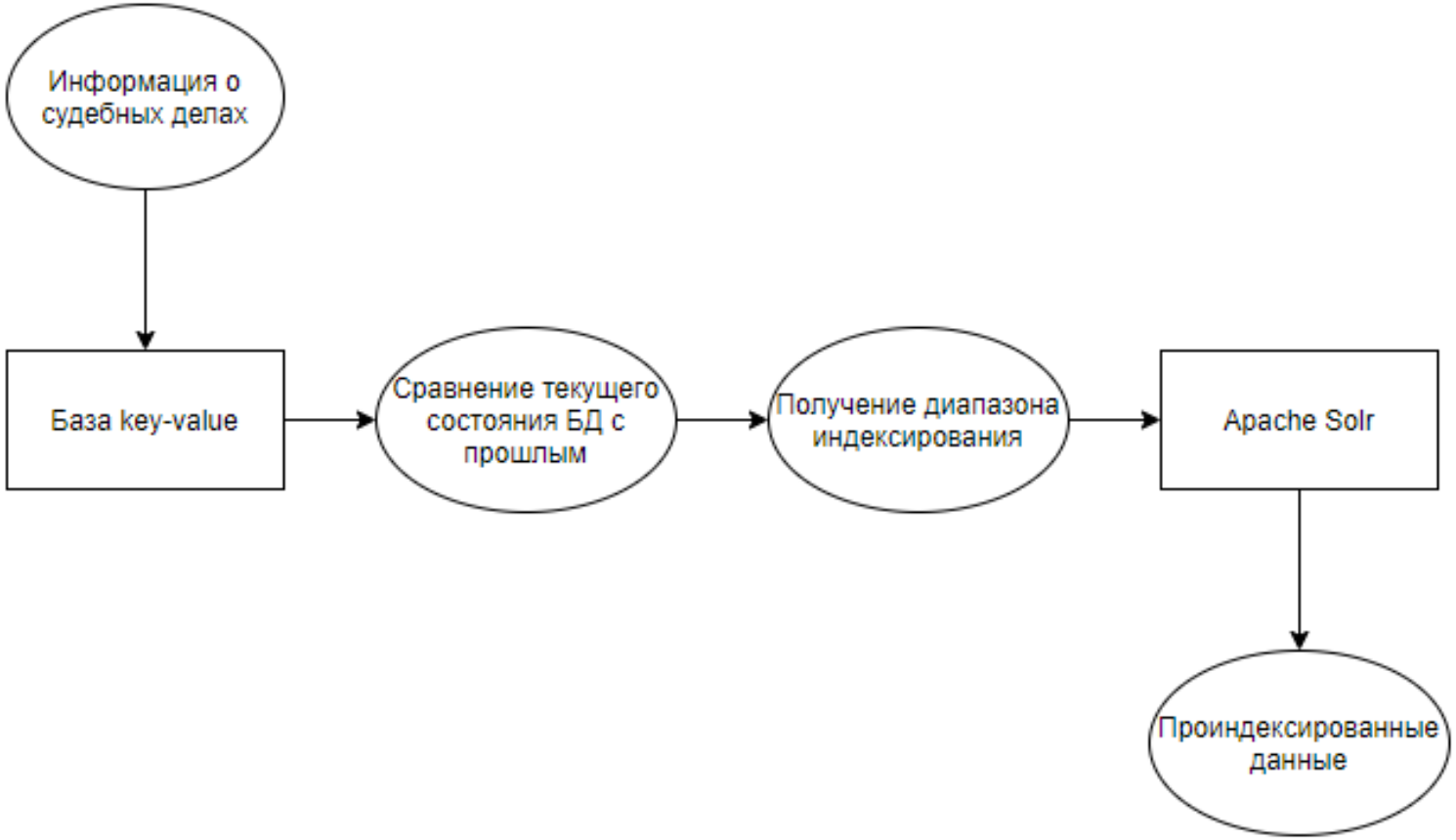
Цель

Ускорение и оптимизация полнотекстового поиска по крупной распределённой базе текстов путём итерационной загрузки и индексирования в системе Apache Solr.

Задачи

- Проектирование сервис-ориентированной архитектуры системы
- Проектирование базы данных для отслеживания хода индексации
- Конфигурирование системы Apache Solr

Анализ предметной области
Объект автоматизации



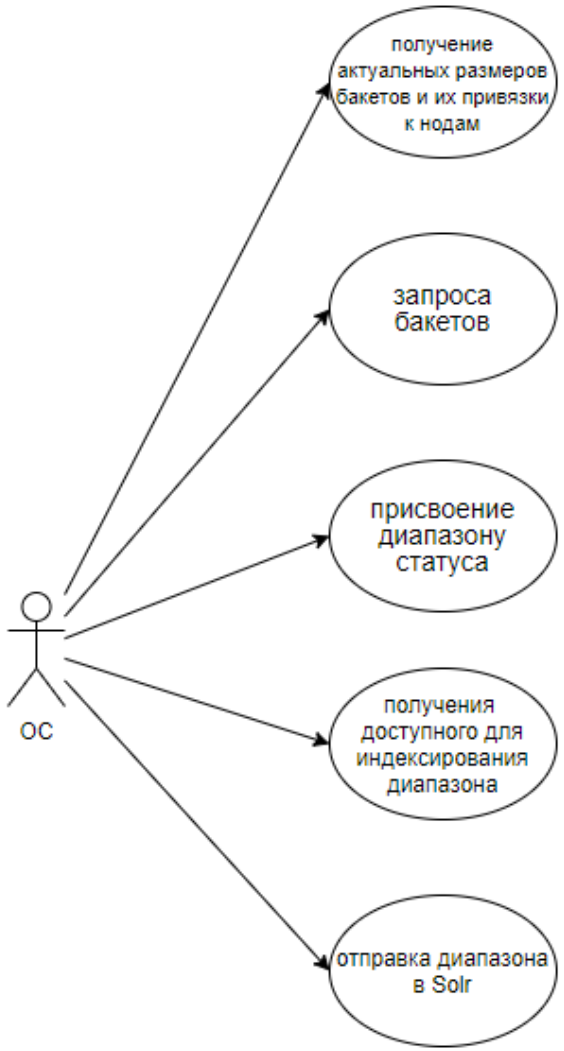
Анализ конкурентов

Система реализует специфическую задачу и создана для обработки данных исключительно «Ирбис Аналитики», поэтому не имеет конкурентов.

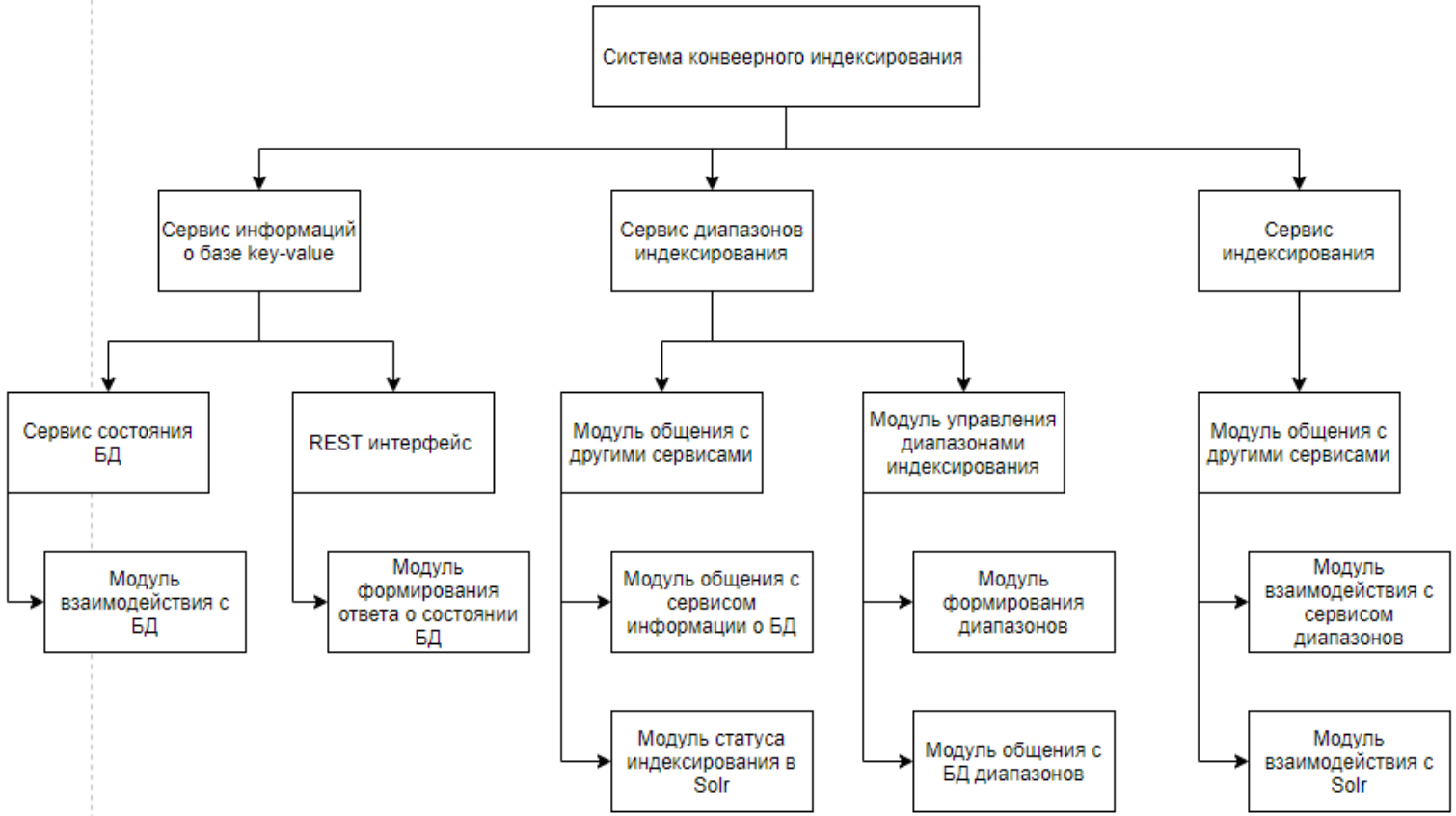
Функции системы

- Отслеживание текущего состояния БД key-value
- Фильтрация и сборка данных из БД
- Формирование диапазонов индексирования
- Отправка данных на индексирование в Apache Solr
- Отслеживание состояния индексирования

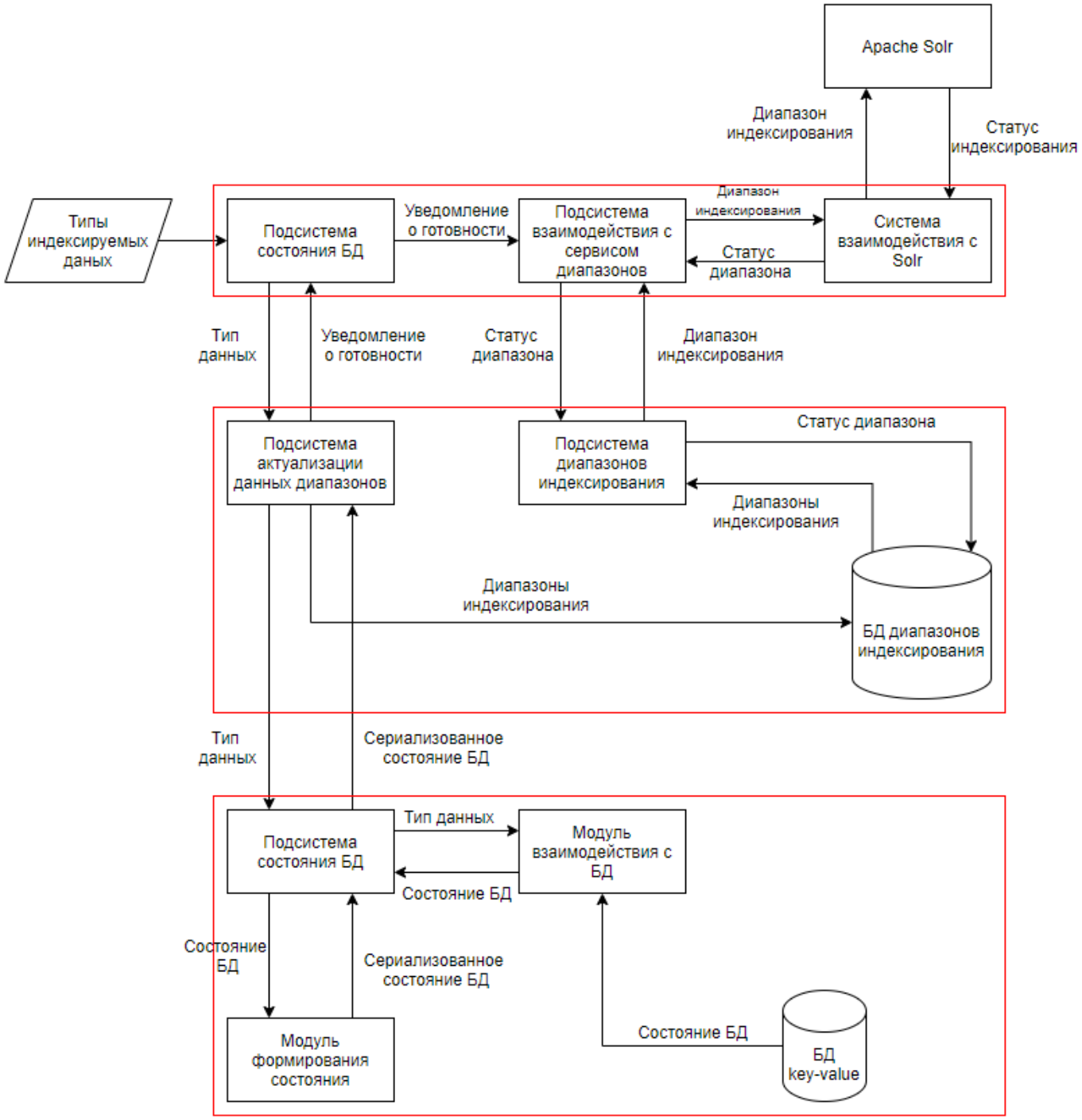
Варианты использования



Структурная схема

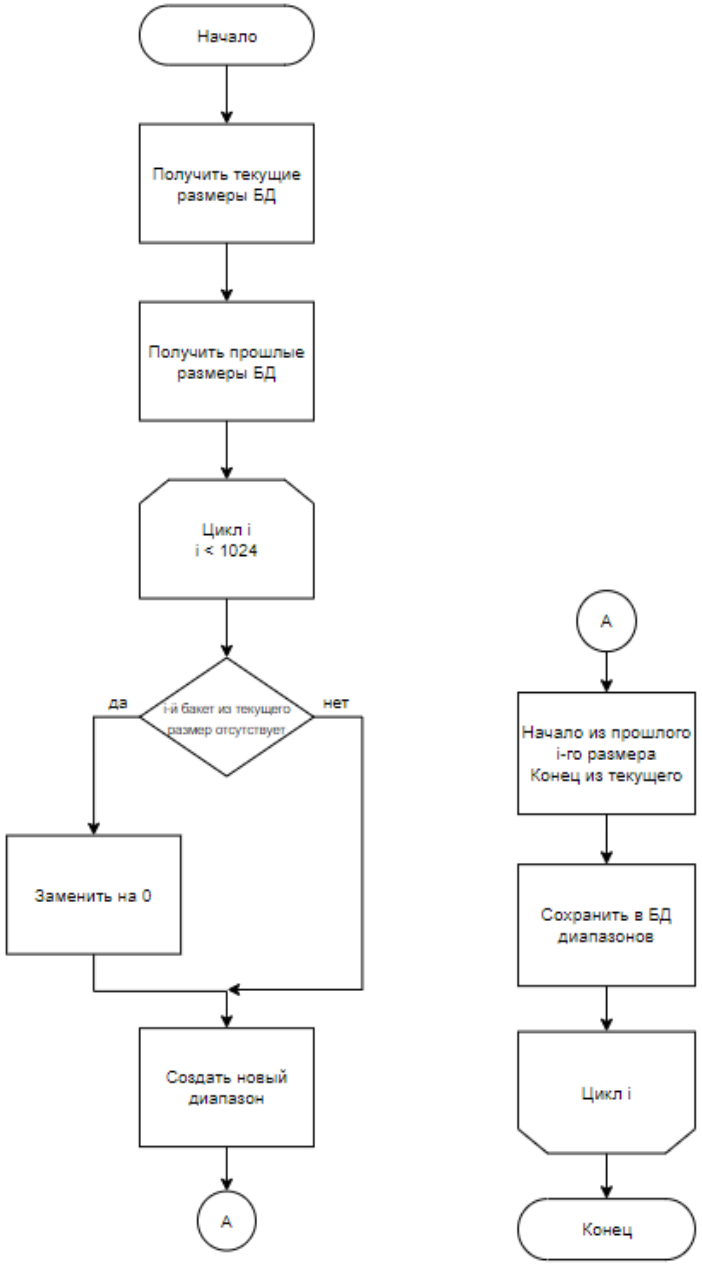


Связи между модулями системы

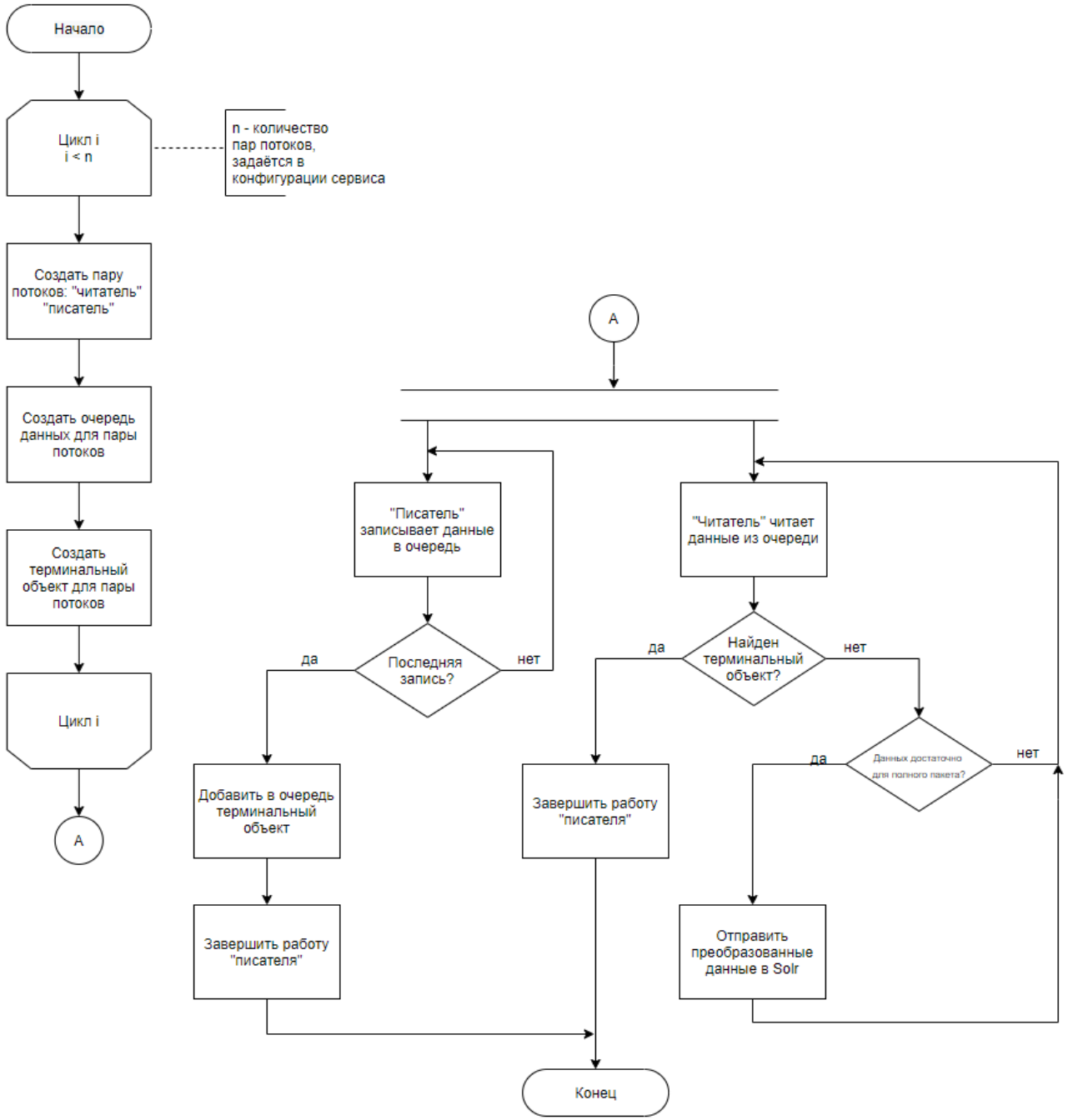


Алгоритм получения информации о базе данных key-value

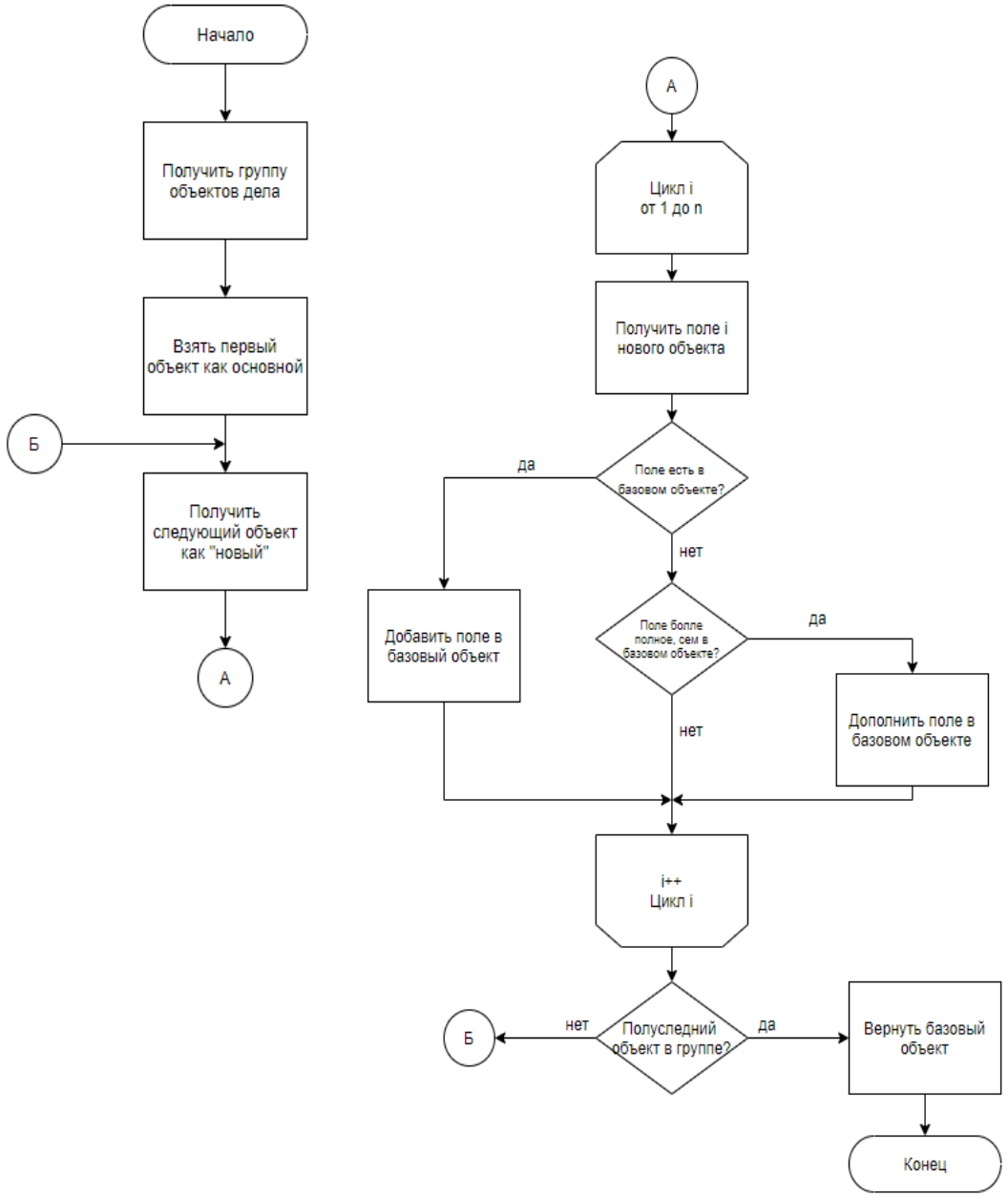
Алгоритм формирования диапазона



Алгоритм отправки данных на индексацию



Алгоритм преобразования группы данных



Алгоритм работы сервиса индексации

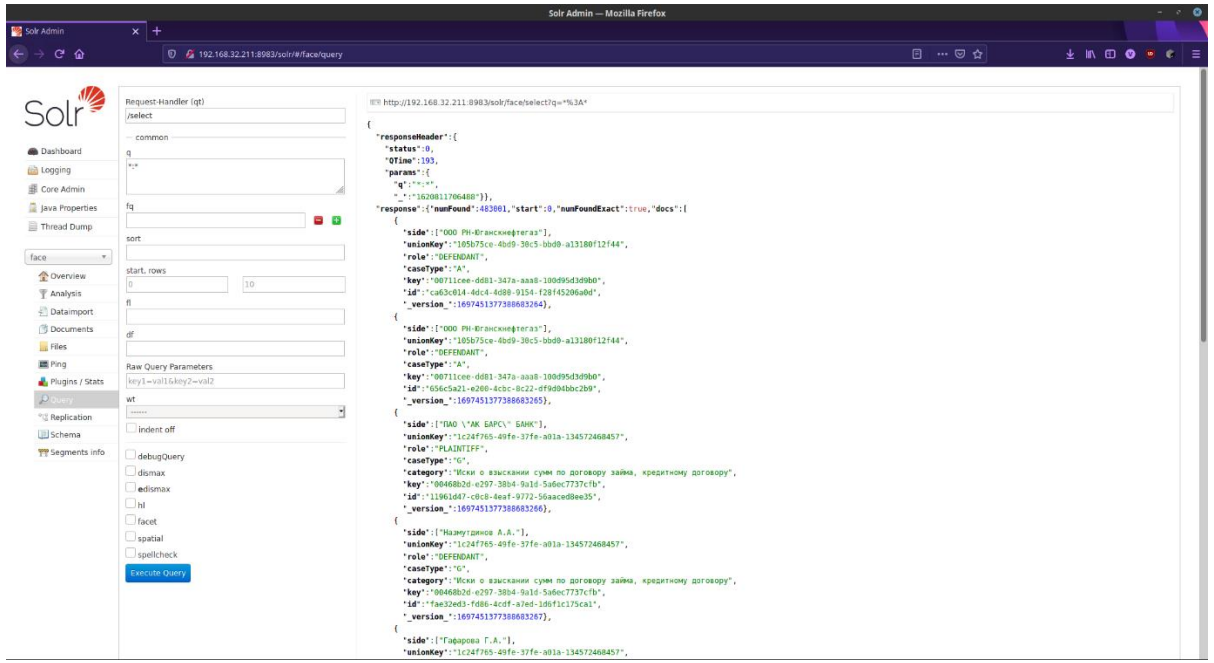
Контрольные примеры

```
GET http://localhost:9002/get_range

HTTP/1.1 200
Content-Type: application/json
Transfer-Encoding: chunked
Date: Thu, 13 May 2021 06:05:48 GMT
Keep-Alive: timeout=60
Connection: keep-alive

{
  "id": 4,
  "bucket_index": 1,
  "node_index": 1,
  "from_order_value": 0,
  "to_order_value": 5
}
```

Объект диапазона



Пример проиндексированных данных

Спасибо за внимание