

Общество с ограниченной ответственностью

«Ред Софт Центр»

Система управления векторными данными

«РЕД Вектор Данных»

Информационный лист

Листов 7

Москва, 2025

Оглавление

Термины и сокращения	3
1. Описание продукта и его характеристик	4
1.1 Общие сведения.....	4
1.2 Совместимость.....	4
2. Состав РЕД Вектор Данных	6
3. Описание производственного цикла с указанием используемых средств	7

Термины и определения

Термин	Определение
Векторный поиск	Поиск ближайших соседей в высокомерных пространствах с использованием метрик косинусного сходства, евклидова расстояния и скалярного произведения.
Embeddings	Семантические представления неструктурированных данных в виде векторов, созданных моделями ИИ.
Payload	Метаданные, ассоциированные с векторами, поддерживающие гибкую фильтрацию запросов.
HNSW	Алгоритм Hierarchical Navigable Small World для приближенного поиска ближайших соседей.
SDK	Набор инструментов для разработки программного обеспечения, включающий библиотеки, API, документацию и примеры кода для создания приложений под конкретную платформу или продукт.
Гибридный поиск	Комбинация векторного поиска по смыслу с фильтрацией по атрибутам payload.
RAG	Retrieval-Augmented Generation — механизм памяти для LLM, обеспечивающий доступ к релевантным данным из базы знаний.
Коллекции	Логические группы векторов с одинаковыми параметрами
Мульти-векторы	Поддержка нескольких векторов для одного объекта.
Разреженные векторы	Векторы для представления данных с большим количеством нулевых значений.
Квантование	Сжатие векторов для экономии памяти и ускорения поиска.
RESTful API	HTTP API для взаимодействия с платформой
gRPC	Высокопроизводительный протокол для клиентских SDK.
Снапшоты	Механизм резервного копирования и восстановления данных коллекций.
Мультитенантность	Изоляция коллекций и данных для разных пользователей в одном кластере.
WSL2	Windows Subsystem for Linux 2 — среда для запуска Linux-контейнеров на Windows.
x86	Intel 80x86 — архитектура процессора и одноимённый набор команд, впервые реализованные в процессорах компании Intel.
x64	64-битная версия (изначально — расширение) архитектуры x86, разработанная компанией AMD и представленная в 2000 году, позволяющая выполнять программы в 64-разрядном режиме.

1 Описание продукта и его характеристик

1.1 Общие сведения

Система управления векторными данными РЕД Вектор Данных — это высокопроизводительное и масштабируемое решение для векторного поиска и управления семантическими представлениями данных (embeddings). Продукт предназначен для разработчиков и компаний, которым необходимо эффективно работать с неструктурированными данными (текст, изображения, аудио, видео), используя современные модели искусственного интеллекта. Мы предоставляем надежную, готовую к производственной эксплуатации платформу с расширенной поддержкой, мониторингом и инструментами для упрощения развертывания.

Функциональные возможности:

- **Высокоскоростной векторный поиск:** Молниеносный поиск ближайших соседей (ANN) даже в сверхвысокомерных пространствах (до 65535 измерений). Поддержка метрик косинусного сходства, евклидова расстояния и скалярного произведения.
- **Гибкая фильтрация:** Уникальная комбинация векторного поиска с условиями на основе payload (метаданных). Позволяет выполнять запросы типа «Найди похожие изображения красных машин, выпущенных после 2020 года».
- **Масштабируемость и отказоустойчивость:** Горизонтальное масштабирование, репликация данных, отказоустойчивые кластерные конфигурации. Идеально для растущих объемов данных и высоких нагрузок.
- **Полноценное API и богатый набор SDK:** RESTful API, gRPC, а также готовые клиентские SDK для Python, Go, Java, Rust и других языков. Легкая интеграция в существующие ML-пайплайны.
- **Управление через Web-интерфейс:** Встроенный административный интерфейс для мониторинга кластера, управления коллекциями, инспекции данных и выполнения тестовых запросов.
- **Гибкие варианты развертывания:** Предлагаем развертывание в вашем облаке (AWS, Yandex Cloud, Selectel), on-premise или в виде управляемого сервиса (Red Vector Data Cloud).
- **Расширенная поддержка и консалтинг:** Гарантированное время реакции, помощь в оптимизации, проектировании архитектуры и интеграции с вашими ML-моделями (OpenAI, Cohere, Hugging Face и др.).

РЕД Вектор Данных может использоваться для решения следующих задач:

1. **Семантический и гибридный поиск:** Умный поиск по документам, товарам, новостям, где запрос понимается по смыслу, а не только по ключевым словам.
2. **Системы рекомендаций:** Рекомендация контента, товаров или услуг на основе векторных представлений предпочтений пользователя и атрибутов объектов.

3. **Поиск по изображениям и видео:** Поиск визуально похожих или тематически связанных изображений (например, для модного ритейла или медиаархивов).
4. **Поиск дубликатов и модерация:** Обнаружение почти идентичного контента (текст, изображения) для борьбы с плагиатом, фродом или NSFW-материалами.
5. **Кластеризация и анализ данных:** Группировка объектов (например, отзывов клиентов, профилей) в векторном пространстве для выявления паттернов и тем.
6. **Память для LLM (RAG):** Критически важный компонент для современных чат-ботов. Позволяет LLM получать релевантные, актуальные данные из вашей базы знаний для формирования точных и обоснованных ответов.

Преимущества перед традиционными базами данных

- **Понимание смысла:** Работает с семантикой данных, а не с точным совпадением строк.
- **Гибкость запросов:** Мощный гибридный поиск, объединяющий смысловое сходство и атрибутные фильтры.
- **Производительность на больших данных:** Специализированная архитектура, созданная именно для векторных операций, в отличие от SQL/NoSQL баз с векторными расширениями.
- **Надежность и контроль:** Платформа корпоративного уровня с полным контролем над развертыванием и конфигурацией.

1.2 Совместимость

РЕД Вектор Данных совместим со следующими программными продуктами:

1. Для клиентов на различных языках:

- Python: Python 3.7 и выше.
- Node.js: Node.js 14 и выше.
- Go: Go 1.18 и выше.
- Rust: Rust 1.65 и выше.
- Java: Java 8 и выше.
- .NET: .NET 6 и выше.
- PHP: PHP 8.0 и выше.

2. Полнофункционально работает на ОС семейства Linux (Ubuntu 20.04+, Debian 11+, CentOS 8+, RHEL 8+).

На macOS (11.0 Big Sur и выше), Windows (10 и выше) через Docker или WSL2 в режиме ограниченной функциональности. РЕД ВЕКТОР ДАННЫХ также

совместим с отечественными ОС: «Ред ОС» (версия 7.3 и выше), «Альт Линукс» (10.1 и выше), включенными в Реестр отечественного ПО.

3. **Поддерживаемые архитектуры процессора:** x86_64, ARM64 (aarch64).
4. **Для развертывания в Kubernetes кластере:** kubectl версии 1.21 и выше, Helm 3.8 и выше.
5. **Поддерживаемые контейнеризационные платформы:** Docker 20.10 и выше, Podman 4.0 и выше.

2. Состав РЕД Вектор Данных

Параметр	Описание
Архитектура ядра	Высокооптимизированное ядро, написанное на языке Rust для максимальной производительности и безопасности
Поддерживаемые типы векторов	Одиночные, мульти-векторы, разреженные векторы
Поисковые алгоритмы	HNSW, Exact Search, Scann
Хранение метаданных	JSON-объекты с гибкой схемой, индексируемые поля
Типы данных метаполей	Integer, Float, Bool, String, Geo, Text (с токенизацией)
Совместимость с ML-экосистемой	Полная интеграция с PyTorch, TensorFlow, OpenAI API, SentenceTransformers, Hugging Face и др.
Форматы импорта/экспорта	JSONL, NDJSON, интеграция с Apache Spark и другими ETL-инструментами

2 Описание производственного цикла с указанием используемых средств

Разработка программного продукта РЕД Вектор Данных ведется с использованием свободных сред разработки, таких как: IntelliJ Idea Community Edition и других. Для управления версиями используется система управления версиями Git — свободная распределённая система управления версиями. Сборка программных продуктов производится на специально выделенных для этого виртуальных серверах, доступ к которым имеет ограниченное число лиц, уполномоченных для работы с ними.

Для сборки используется утилита Cargo рекомендуемая версия не ниже 1.91.1.

Разработка ведется на языке программирования Rust версии 8 и 11.

Общие стадии технологического процесса производства продукции:

- 1 Входной контроль комплектующих элементов, используемых для продукции.
- 2 Запись на оптический носитель продукции.
- 3 Контроль качества записи на компакт-диске.
- 4 Контроль соответствия записанного образца продукции эталонному образцу.
- 5 Заполнение формуляра, маркировка и упаковка.

Примечание: При несоответствии качества или несоответствии эталонному образцу продукция бракуется и изымается из поставки.