

Постановка на разработку Smart Platform

Smart Platform

1. Общие сведения

1.1 Назначение

Разработка B2B сервиса для управления корпоративными знаниями с возможностями автоматической обработки данных на основе RAG (Retrieval-Augmented Generation) архитектуры.

1.2 Обзор архитектуры системы

Система состоит из следующих основных компонентов:

- Фронтенд-приложение (Vue.js)
- Бэкенд-сервисы (Python)
- RAG конвейер
- Векторная база данных
- Сервис аутентификации и авторизации
- Система автоматической валидации (RAGAS)

2. Функциональные требования

2.1 Управление доменами знаний

- Создание и управление доменами знаний
- Настройки конфигурации для каждого домена
- Настраиваемые поля метаданных
- Правила классификации документов
- Интеграция с внешними источниками данных

2.2 Конвейер обработки документов

Поддержка различных форматов файлов:

- PDF
- DOCX
- PPTX
- Excel файлы
- Веб-контент
- Автоматическое извлечение текста
- Векторизация документов
- Извлечение метаданных
- Система автоматической разметки

2.3 Реализация RAG конвейера

- Модуль понимания запросов
- Система извлечения контекста
- Движок поиска документов
- Сервис реранжирования
- Оценка семантического сходства
- Оценка релевантности
- Настраиваемые правила ранжирования
- Система генерации ответов

2.4 Администрирование и управление пользователями

- Ролевая модель доступа (RBAC):
 - Суперадминистратор
 - Администратор домена
 - Обычный пользователь
- Мониторинг активности пользователей
- Управление конфигурацией системы

- Управление API-ключами
- Аналитика использования и отчетность

2.5 Система автоматической валидации

- Автоматическое тестирование RAG конвейера
- Мониторинг метрик качества:
 - Релевантность ответов
 - Точность ответов

3. Технические требования

3.1 Фронтенд-приложение

- Фреймворк: Vue.js 3
- Компоненты:
 - Интерфейс загрузки документов
 - Управление доменами знаний
 - Интерфейс взаимодействия с системой
 - Панель администрирования
 - Панель аналитики
- Особенности:
 - Адаптивный дизайн
 - Обновления в реальном времени
 - Предпросмотр документов
 - Интерактивная визуализация

3.2 Бэкенд-сервисы

- Язык: Python 3.11+
- Фреймворк: FastAPI
- Основные сервисы:
 - Сервис обработки документов
 - Сервис RAG конвейера
 - Сервис аутентификации
 - Сервис администрирования
 - Сервис аналитики

3.3 Архитектура баз данных

- Векторное хранилище:
 - Milvus
 - Индексированные эмбединги документов
 - Хранение метаданных
- SQL база данных:
 - Управление пользователями
 - Конфигурация системы
 - Аналитические данные
 - Метаданные документов

3.4 Требования к интеграции

- Поддержка внешних коннекторов:
 - Confluence
 - Google Drive
 - Jira
 - CRM системы
 - Пользовательские коннекторы через API
- Интеграция аутентификации:
 - OAuth 2.0
 - SAML
 - Active Directory

4. Нефункциональные требования

4.1 Производительность

- Время ответа на запрос: < 2 секунд

- Время обработки документов: < 5 минут на 100 страниц
- Поддержка одновременных пользователей: 100+
- Доступность системы: 99.9%

4.2 Безопасность

- Шифрование данных в покое и при передаче
- Регулярные аудиты безопасности
- Логирование контроля доступа
- Соответствие отраслевым стандартам:
 - GDPR
 - SOC 2

4.3 Масштабируемость

- Возможность горизонтального масштабирования
- Балансировка нагрузки
- Система кэширования
- Оптимизация ресурсов

5. Разработка и развертывание

5.1 Среда разработки

- Система контроля версий: Git
- CI/CD конвейер
- Тестовое окружение:
 - Модульные тесты
 - Интеграционные тесты
 - Нагрузочное тестирование
 - Тестирование безопасности

5.2 Развертывание

- Контейнеризация: Docker
- Оркестрация: Kubernetes
- Конфигурация окружений:
 - Разработка
 - Тестирование
 - Продакшн

6. Требования к документации

- Документация API
- Документация по архитектуре системы
- Руководство пользователя
- Руководство администратора
- Руководство по интеграции

7. Критерии успешности

- Успешная обработка поддерживаемых форматов документов
- Точная классификация документов
- Релевантные результаты поиска
- Соответствие метрикам производительности
- Выполнение требований безопасности
- Успешное прохождение автовалидации