

INFATON Chat

Техническая архитектура

Версия 17.11.419-registry.39 · Редакция 15 октября 2026 года

INFATON Chat - веб-приложение с серверным ядром и отдельными сервисами обработки данных. Пользователь работает в браузере; ядро управляет сеансами, прикладными сценариями, доступом, хранением и обменом с подключёнными системами. Документ описывает серверную архитектуру версии 17.11.419-registry.39; конкретный состав компонентов определяется паспортом и манифестом поставки.

Структура системы

Архитектура версии .39 сочетает прикладное ядро на Go и отдельные сервисы обработки. В принятом манифесте поставки учтены образы приложения, PostgreSQL, OCR, аналитики, эмбедингов и распознавания речи; локальные веса BGE-M3 включены в пакет. Наличие сервиса или весов не означает, что соответствующий профиль включён в настройках. Разделение на контейнеры помогает управлять составом и жизненным циклом сервисов, но само по себе не запрещает сетевой доступ. Браузер обращается к ядру по HTTPS через настроенный прокси. Ядро использует PostgreSQL и внутренние интерфейсы сервисов. Основная языковая модель и 1C подключаются отдельно.

Компонент	Назначение и связь
Браузер	Интерфейс JavaScript, формы, история, представление результатов
HTTPS-прокси	Вход пользователей, сертификат и маршрутизация к приложению
Приложение на Go	Авторизация, прикладная логика, API, управление обменом
PostgreSQL с pgvector	Пользователи, настройки, история, прикладные данные и поиск
OCR	Извлечение текста и данных из PDF и изображений
Аналитика	Расчёт показателей запасов и сценарии по данным 1C; включается профилем analytics
Эмбединги	Построение векторов для базы знаний; веса BGE-M3 находятся в пакете, доступ зависит от профиля rag и конфигурации
Распознавание речи	Преобразование аудио в текст; профиль voice
Основная модель	Отдельный OpenAI-совместимый endpoint для диалога и сценариев
MCP-мост и 1C	Расширение MCP 6.2.8 связывает приложение с опубликованным мостом и целевой базой; сама платформа 1C не входит и лицензируется отдельно

Docker Compose связывает сервисы внутренней сетью и именованными томами. Приложение по умолчанию публикуется на loorback-адресе сервера; доступ пользователям обычно предоставляет настроенный HTTPS-прокси. Основная модель, 1C и прокси размещаются отдельно от контейнеров приложения, если паспорт поставки не устанавливает иное. Права доступа и сетевые ограничения для этих узлов задаются в схеме внедрения.

Реализация и зависимости

В состав входят обработка документов (OCR), расчёт показателей запасов, преобразование речи в текст, эмбединги для базы знаний и обработка IFC. OCR, аналитика, эмбединги и распознавание речи поставляются отдельными сервисами; IFC обрабатывается в ядре приложения. Их доступность зависит от профиля, настроек, прав и входных материалов.

Приложение использует Go и JavaScript; сервисы обработки включают Python и сторонние библиотеки, а распознавание речи содержит нативные компоненты. Точный состав образов и файлов задаётся подписанным манифестом конкретной поставки. В архиве размещены LICENSE.txt и уведомления по серверным, нативным и модельным компонентам. Сторонние части используются по условиям соответствующих правообладателей и не являются авторской разработкой ООО «ИНФАТОН». Платформа 1C, основная языковая модель и внешний HTTPS-прокси не включены в серверный комплект INFATON Chat.

Потоки данных

Запрос пользователя поступает в ядро с контекстом его сеанса. Ядро определяет доступные источники и направляет необходимые данные выбранной модели. Для операций 1C используется отдельное подключение к MCP-мосту; ограничения целевой базы должны действовать вместе с прикладными проверками. Результат сохраняется и возвращается в интерфейс. Для операций с подтверждением сервер сохраняет решение и связывает предложение с конкретным подключением. Повторное подтверждение и восстановление браузерного сеанса не должны выполнять тот же эффект заново. Неизвестный результат обмена требует проверки в целевой системе; его нельзя обозначать как успешную запись.

Файл поступает в приложение и предусмотренный сервис обработки. Для поиска документ проходит извлечение текста и построение эмбедингов. Ответ пользователю должен соответствовать его области доступа. Голосовая запись передаётся сервису речи; полученный текст используется как вход диалога. Конкретный набор передаваемых данных зависит от функции и настроек.

Внешняя анкета или тендер предоставляют ограниченный интерфейс по ссылке. Ссылка не создаёт внутреннюю учётную запись. Доступ определяется обработчиком формы и состоянием ссылки. Перед выдачей и после отзыва проверяют фактическую видимость сведений в отдельном сеансе.

Доступ и секреты

Права проверяются на сервере; скрытие кнопки в браузере не заменяет серверную проверку. Общие документы базы знаний открываются обычным пользователям по явно заданным допускам групп, пользователей или разрешению для всех вошедших. Область конкретной базы 1C и доступ к разделу учитываются отдельно. Учётная запись владельца обладает расширенными полномочиями и не подходит для оценки прав обычного пользователя. Для каждой базы 1C настраиваются соответствующие подключения и права.

ERPTWIN_SECRETS_KEY применяется для защиты сохраняемых секретов подключений. Его потеря может сделать ранее сохранённые настройки непригодными; при восстановлении нужен тот же ключ. Ключ входит в полную резервную копию. Сам архив не зашифрован, поэтому доступ к нему и носителю ограничивают административными правами.

Закрытый контур и внешние интеграции

Модель может быть локальной или удалённой. При удалённом размещении текст запросов и контекст передаются выбранному endpoint. Для работы без внешней модели все необходимые веса, сервисы и зависимости должны быть доступны внутри инфраструктуры заказчика. Контейнеризация приложения сама по себе не доказывает отсутствие исходящих соединений; сетевую схему проверяет и ограничивает администратор.

SMTP, мессенджеры, внешние проверки и другие интеграции включаются отдельно. До их активации определяют получателей, состав данных и сетевые правила. Наличие настроек маскирования не заменяет проверку всей схемы обработки и организационных мер заказчика.

Состояние и восстановление

Постоянное состояние распределено между PostgreSQL, файловыми томами и конфигурацией, включая ACL и ключ шифрования сохранённых секретов. Штатный backup-box.ru формирует полную копию выбранной установки. Восстановление выполняется в новую папку и новый Compose-проект с точными исходными образами. Архив содержит секреты и не зашифрован; он не заменяет исходные образы и файлы моделей. Внешние 1С, модель, прокси и ОС резервируются отдельно. Проверка /healthz относится к процессу приложения, /readyz - к его готовности работать с БД; для остальных функций нужны самостоятельные контрольные действия.

Серверная поставка предназначена для Linux x86_64 (amd64) с Docker Engine и Docker Compose v2. Совместимость с конкретным дистрибутивом и его версией проверяют перед установкой. Доступность, отказоустойчивость и производительность определяются размещением и нагрузкой. Единственная установка Compose не является кластером высокой доступности. Требования конкретного предприятия проверяются на его согласованной конфигурации.