DeepDive

Архитектура и безопасность

Версия 0.1.0 · 3 ноября 2026 г.

О DeepDive

DeepDive — ИИ-ассистент для анализа данных компании. Сотрудник задаёт вопрос на обычном языке — например, «почему упали продажи в Москве?» — и получает ответ по данным своей компании: выводы, таблицы и графики. Ассистент работает только с подключёнными данными и файлами пользователя, интернет для ответов не используется.

Для кого

- **Бизнес-пользователи** — руководители, аналитики, сотрудники подразделений.
- **Администратор организации** — источники данных, участники, бюджет, правила ассистента.
- **ИТ-служба и служба безопасности заказчика** — развёртывание, сопровождение, аудит.

Как устроено

Данные из источников — баз данных, витрин, учётных систем, файлов — приводятся к единой модели с одинаковыми показателями и справочниками.

Ассистент отвечает на вопросы по этой модели, графики из ответов собираются в дашборды, а автопилоты по расписанию проверяют данные и присылают сигналы о рисках и возможностях.

Платформа разворачивается как SaaS, on-premise в контуре заказчика или в частном облаке. Продукт везде один и тот же, различаются место хранения данных и то, кто обновляет платформу.

Состав документации

- Руководство пользователя — для пользователей и администраторов организации.
- Установка и администрирование — для ИТ-службы заказчика.
- **Архитектура и безопасность (этот документ)** — для архитекторов и службы ИБ.

Поддержка

Платформу предоставляет DeepDive. Вопросы о работе платформы — support@godeepdive.ru.

Содержание

Архитектура и безопасность

Обзор 4
Архитектура 5
Обработка запроса 11
Данные 13
Скрытие данных от модели 15
MCP и интеграции 17
Доступ и аудит 19
Сеть и порты 20
Ответы на типовые вопросы ИБ 21
Потоки данных 24
Модели и маршруты 29
Хранение и удаление данных 32
Изоляция тенантов 35
Аутентификация и роли 38
Журналирование и аудит 41
Сетевой периметр 43
SaaS: обработка данных и SLA 45

Обзор

Документ — для архитекторов и службы информационной безопасности заказчика на этапе знакомства с платформой. Основная часть — восемь коротких глав: каждая открывается схемой или таблицей и читается за пару минут. Подробности, на которые главы ссылаются, собраны в приложениях. Сведения сверены с исходным кодом и дистрибутивом платформы; где функция не реализована или ограничена, это сказано прямо.

Главы: архитектура (стр. 5), обработка запроса (стр. 11), данные (стр. 13), скрытие данных от модели (стр. 15), MCP и интеграции (стр. 17), доступ и аудит (стр. 19), сеть и порты (стр. 20).

Что решает заказчик

В установке, которую эксплуатирует сам заказчик, решения о безопасности принимает он: какую модель использовать и разрешать ли внешний API; включать ли трассировку; какие источники подключить и с какими правами; сроки хранения журналов, файлов и копий; чей Keycloak использовать для входа; какие исходящие соединения разрешить. Администратор организации дополнительно решает, какие данные скрывать от модели.

Термины

Термин	Значение
Установка	Одна инсталляция платформы: сервисы, хранилища, Keycloak
Организация, тенант	Изолированное пространство заказчика или подразделения в установке: участники, настройки, источники, история диалогов
Оператор	Тот, кто эксплуатирует установку: в SaaS — компания-оператор, в on-premise — ИТ заказчика
Провайдер модели	Компания или система, предоставляющая API языковой модели
Шлюз моделей	Необязательный прокси между платформой и провайдером: маршрутизация, учёт и лимиты расходов

Архитектура

Платформа одна во всех вариантах развёртывания: набор компонентов тот же, меняется, кому принадлежит контур, в котором они работают, и куда идут соединения. Ниже — состав установки и границы доверия по вариантам.

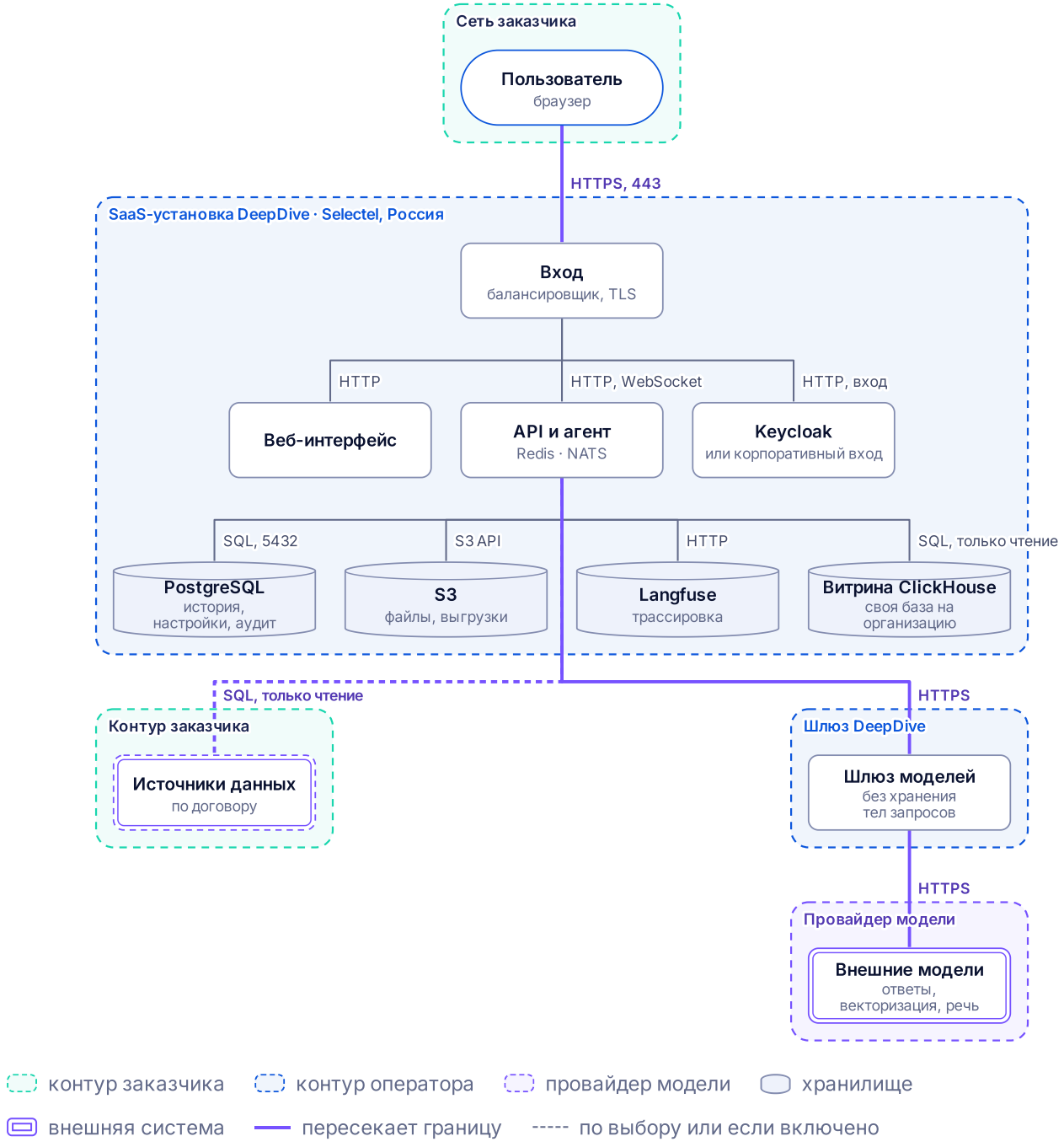


Рис. 1. Архитектура: SaaS. Установка общая для нескольких заказчиков и принадлежит DeepDive; данные каждой организации — в своём тенанте. Запросы к модели идут через шлюз DeepDive к провайдеру модели.

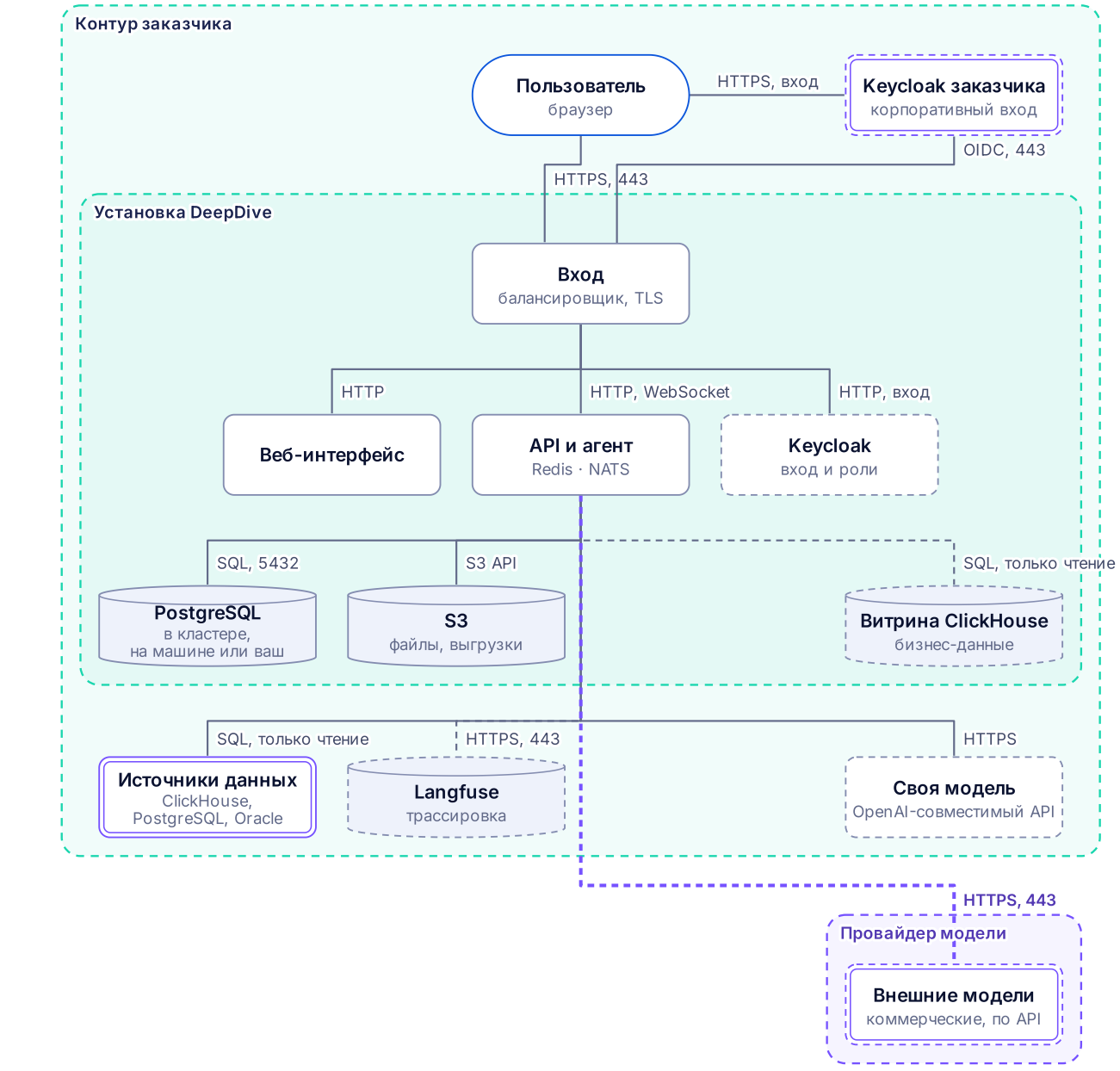


Рис. 2. Архитектура: On-premise. Вся платформа — в контуре заказчика, ставится офлайн-дистрибутивом и связи с вендором не имеет. Контур покидают только запросы к модели, если она внешняя.

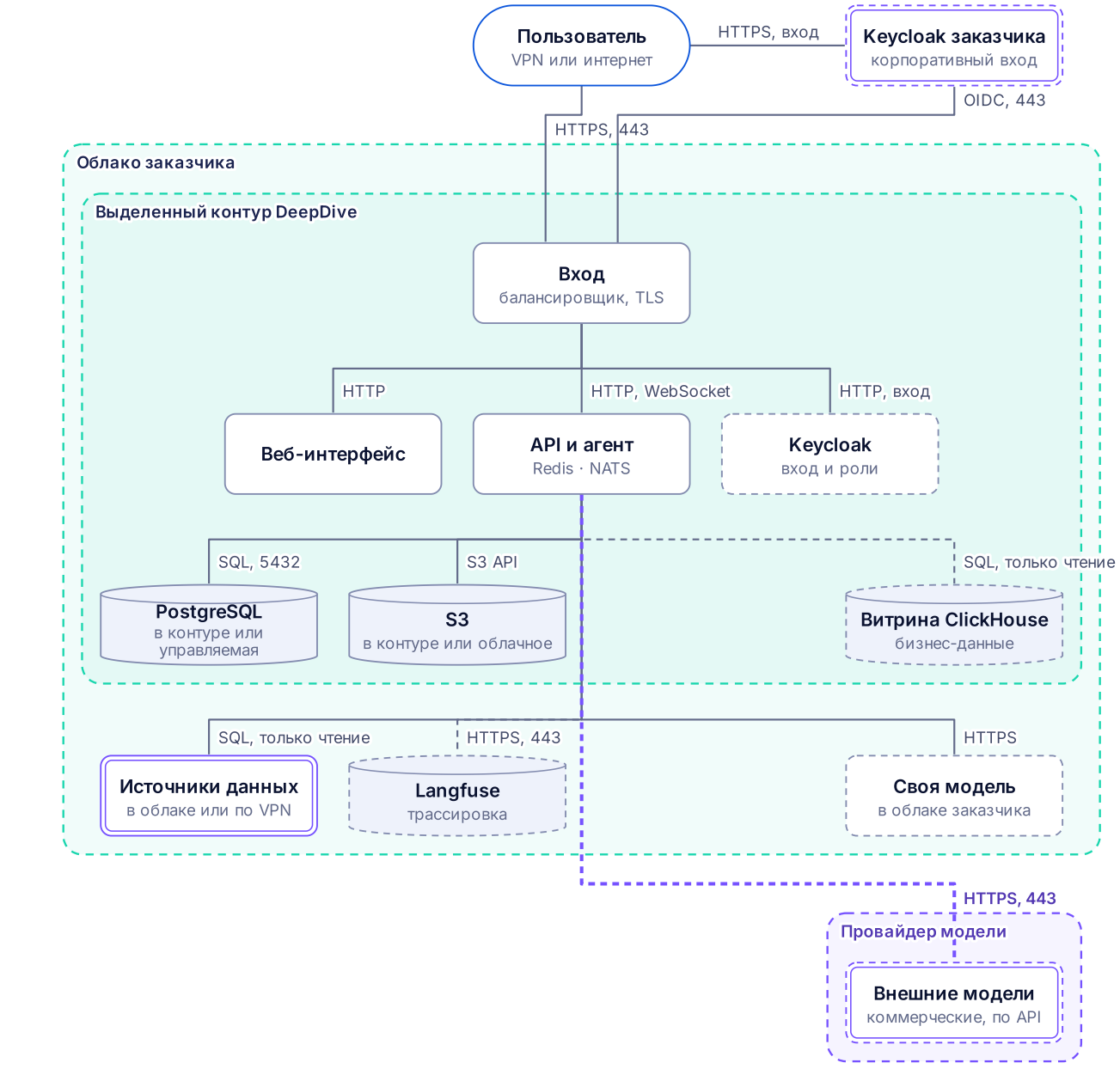


Рис. 3. Архитектура: Частное облако. Выделенная установка для одного заказчика в облаке, которое он выбрал; ресурсы ни с кем не делятся. Как и в on-premise, контур покидают только запросы к внешней модели.

Компоненты

Состав — по офлайн-дистрибутиву: всё, что в таблице, ставится из комплекта или подключается вместо компонента комплекта. Коннекторы MCP (почта, мессенджер) и шлюз моделей в комплект не входят.

Компонент	Назначение и данные	SaaS	On-premise	Частное облако
Вход (ingress) и TLS	Публикует домены по HTTPS, терминирует TLS. Данные: нет	Установка DeepDive	Envoy Gateway из комплекта или ingress заказчика	Как on-premise или балансировщик облака

Компонент	Назначение и данные	SaaS	On-premise	Частное облако
oauth2-proxy	Единственная точка входа: проверяет сессию в Keycloak. Данные: сессии входа — в Redis	Установка DeepDive	Установка	Установка
Keycloak	Вход, организации и роли пользователей. Данные: учётные записи, группы — в своей базе PostgreSQL	Установка DeepDive; корпоративный вход по согласованию	Из комплекта, свой realm на Keycloak заказчика или Keycloak заказчика только для входа	Как on-premise
Веб-интерфейс	Чат, панели, настройки организации. Данные: нет	Установка DeepDive	Установка	Установка
API и агент	Права, работа ассистента, запросы к источникам и модели. Данные: своих нет, состояние в PostgreSQL и Redis	Установка DeepDive	Установка	Установка
Служебные сервисы	Конфигурация, реестр сервисов, организации. Данные: в PostgreSQL	Установка DeepDive	Установка	Установка
PostgreSQL 16 + pgvector	Основная база платформы. Данные: история диалогов, настройки, журнал изменений, база знаний и индекс	Установка DeepDive	В кластере, на машине из комплекта или у заказчика	Как on-premise или сервис облака
Redis	Кэш платформы и сессии входа. Данные: без долговременных данных	Установка DeepDive	В кластере, на машине из комплекта или у заказчика	Как on-premise
NATS JetStream	Шина событий и потоки ответов ассистента. Данные: сообщения в пути	Установка DeepDive	В кластере или у заказчика	Как on-premise
S3-совместимое хранилище	Файлы пользователей и выгрузки. Данные: файлы, крупные результаты, выгрузки	Установка DeepDive	Garage в кластере, на машине из комплекта или S3 заказчика	Как on-premise или хранилище облака
Витрина ClickHouse	Аналитическое хранилище бизнес-данных. Данные: бизнес-данные	Установка DeepDive, своя база на организацию	Витрина заказчика, пустая из комплекта или без витрины	Как on-premise
Источники данных заказчика	ClickHouse, PostgreSQL, Oracle; учётная запись только для чтения. Данные: у заказчика; платформа хранит каталог схемы	По договору — в контуре заказчика	Контур заказчика	Облако заказчика или ЦОД по VPN
Langfuse	Трассировка работы ассистента. Данные: промпты, ответы, идентификаторы	Langfuse в установке, 30 дней	Langfuse заказчика или выключена	Как on-premise
Шлюз моделей	Маршрутизация к провайдеру, лимиты расходов. Данные: учёт по ключам, без тел запросов	Инфраструктура DeepDive, вне контура заказчика	Нет: платформа обращается к модели напрямую	Как on-premise
Языковая модель и векторизация	Ответы, векторизация, распознавание речи. Данные: по условиям провайдера	Внешние коммерческие модели по API, через шлюз DeepDive	Внешние коммерческие модели по API либо своя модель в контуре	По API или модель в облаке заказчика

Сервисы платформы стартуют по цепочке проверок готовности: служебные сервисы → API → веб-интерфейс. Снаружи установки опубликован только вход: сервисы и хранилища общаются внутри кластера или с машинами установки.

Границы доверия

Граница доверия проходит там, где данные переходят под контроль другой стороны: от заказчика к оператору установки или к провайдеру модели.

Сторона	SaaS	On-premise	Частное облако
Заказчик	Пользователи и их браузеры; источники данных, если подключены из его контура; администрирование своей организации	Вся установка: сервисы, базы, файлы, трассировка, ключи и сертификаты, вход пользователей; модель, если она в контуре	Выделенный контур в своём облачном аккаунте: сервисы, базы, файлы, трассировка, ключи; модель, если она в облаке заказчика
Оператор	Установка DeepDive на площадке Selectel: сервисы, базы, файлы, трассировка, резервные копии; шлюз моделей вне контура заказчика	Нет оператора: вендор поставляет дистрибутив, доступа к установке и данных не получает	Администрирование — по договору: заказчиком, вендором или интегратором; доступ к контуру выдаёт заказчик
Провайдер модели	Внешние коммерческие модели: получает содержимое запросов к модели, тексты для векторизации, аудио голосового ввода	Только если выбрана внешняя модель: получает содержимое запросов к модели и тексты для векторизации	Как on-premise

Что пересекает границу

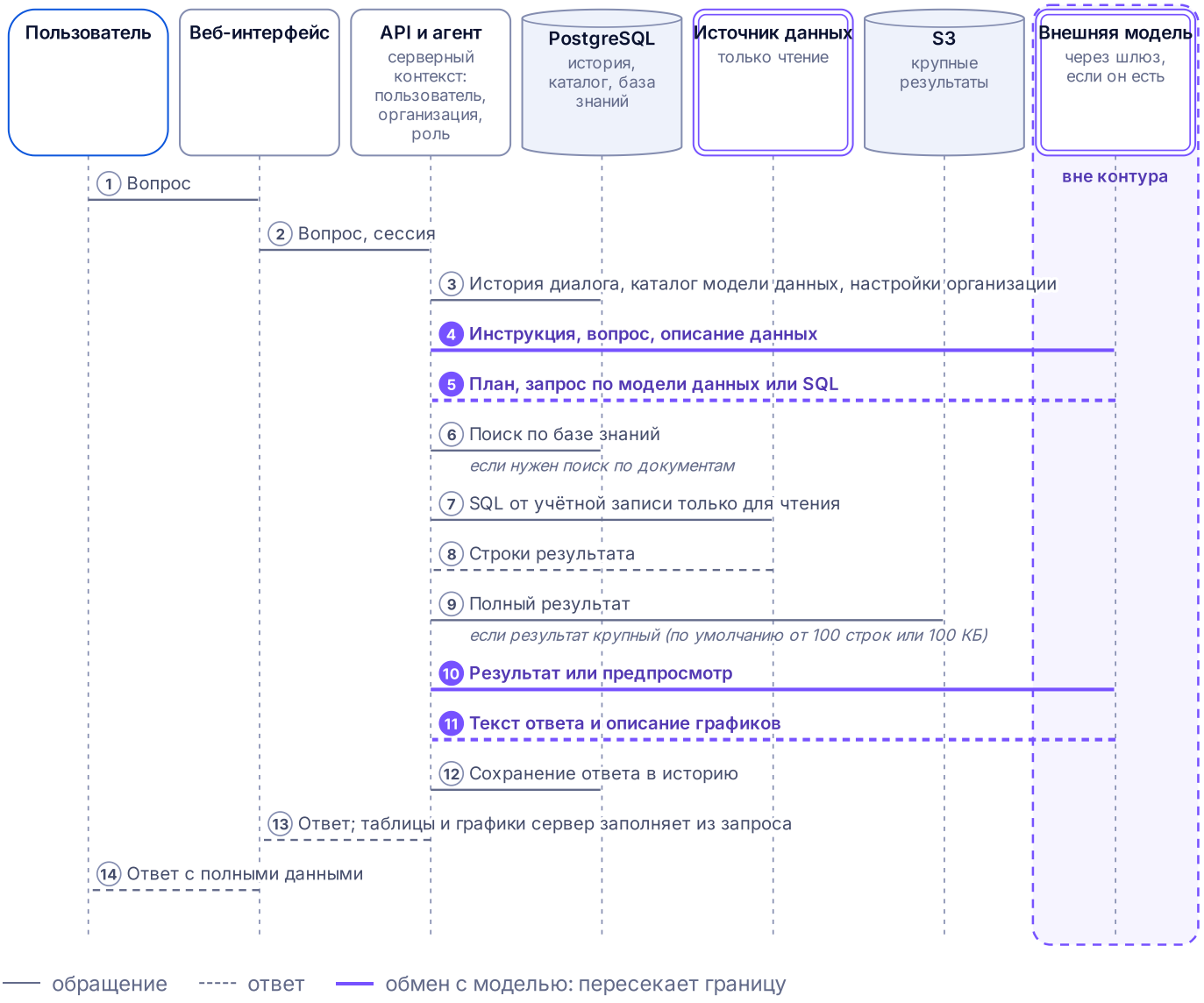
В таблице — все соединения, по которым данные уходят из-под контроля своей стороны. Других нет: платформа не обращается к серверу вендора, не отправляет телеметрию и не проверяет обновления, лицензия проверяется по подписи внутри установки. Коннекторы, MCP-серверы и вебхуки передают наружу то, для чего их подключили, — см. главу «MCP и интеграции» (стр. 17).

Соединение	Что передаётся и зачем	Где
Пользователь → Вход, HTTPS, 443	Работа с платформой: вопросы, ответы, файлы, выгрузки, вход	SaaS
API и агент → Источники данных, SQL, только чтение	Аналитика по данным заказчика: SQL-запросы, обратно — результаты Только по договору	SaaS
API и агент → Шлюз моделей, HTTPS	Ответ ассистента: вопрос с историей, каталог модели данных, результаты в пределах порога, фрагменты базы знаний Векторизация для поиска: тексты документов, описаний схемы и вопроса Распознавание речи: аудио голосового ввода (если включён голосовой ввод)	SaaS
Шлюз моделей → Внешние модели, HTTPS	Ответ ассистента: вопрос с историей, каталог модели данных, результаты в пределах порога, фрагменты базы знаний Векторизация для поиска: тексты документов, описаний схемы и вопроса Распознавание речи: аудио голосового ввода (если включён голосовой ввод)	SaaS
API и агент → Внешние модели, HTTPS, 443	Ответ ассистента: вопрос с историей, каталог модели данных, результаты в пределах порога, фрагменты базы знаний Векторизация для поиска: тексты документов, описаний схемы и вопроса (если модель векторизации внешняя) Только если модель внешняя	On-premise, Частное облако

Сетевые правила для этих соединений — в главе «Сеть и порты» (стр. 20), где лежит каждый класс данных — в главе «Данные» (стр. 13).

Обработка запроса

Ассистент отвечает на вопрос за несколько обращений к модели: план, запрос к данным, разбор результата. Запрос к данным выполняет сервер, а не модель. Модель видит только то, что сервер ей передал, и данные из источников получает в пределах порога.



Обработка одного вопроса. Курсивом под стрелкой — когда шаг выполняется.

№	Шаг	Что получает модель
4	API и агент → Внешняя модель: Инструкция, вопрос, описание данных	Системная инструкция с описанием организации и каталогом модели данных (названия и описания датасетов, метрик, измерений), вопрос и история текущего диалога
6	API и агент → PostgreSQL: Поиск по базе знаний (если нужен поиск по документам)	Найденные фрагменты документов организации

№	Шаг	Что получает модель
10	API и агент → Внешняя модель: Результат или предпросмотр	Строки результата, если он небольшой; иначе предпросмотр (по умолчанию 20 строк) и сводка. При скрытии данных от модели помеченные значения заменены на [redacted]

Порог крупного результата и размер предпросмотра задаются для каждой модели в панели администратора (см. «Установка и администрирование» → «Панель администратора установки»). Таблицы, графики и выгрузки пользователь получает полностью: их заполняет сервер из запроса, модель в этом потоке не участвует.

Не передаются модели никогда: идентификаторы пользователя и организации, имя, адрес электронной почты, роли, учётные данные источников, токены доступа. Они в серверном контексте запроса: права на каждом шаге проверяются по нему, а не по тому, что предложила модель.

Кроме ответа, к провайдеру уходят тексты для векторизации (документы, описания схемы, вопрос при поиске), описание источника с примерами значений — до 50 на колонку — и аудио голосового ввода, если он включён. Подробно — в приложениях «Потоки данных» (стр. 24) и «Модели и маршруты» (стр. 29).

Данные

Класс данных	Где лежит	Срок по умолчанию	Кто имеет доступ
История диалогов: вопросы, ответы, SQL, результаты в объёме, который видела модель	PostgreSQL	Пока диалог не удалят	Автор; администратор организации; администратор установки
Рабочее состояние ассистента	PostgreSQL	Два последних состояния диалога; через 30 дней без активности удаляется	Только сервисы платформы
Настройки организации, модель данных, подключения к источникам	PostgreSQL	Пока их не удалит администратор	Администратор организации; реквизиты общих источников — только администратор установки
База знаний и её векторный индекс	PostgreSQL	Пока документы не удалят	Участники организации, через ассистента
Файлы, прикреплённые к вопросам	S3	30 дней	Пользователь, загрузивший файл, в своей организации
Выгрузки в файл	S3	24 часа; ссылка действует 15 минут	Пользователь, запросивший выгрузку
Крупные результаты запросов, файлы библиотеки	S3	Задаёт оператор политикой хранилища	Участники организации в пределах своих прав
Журнал изменений доступа	PostgreSQL	Бессрочно, переживает удаление организации	Администратор установки
Трассировки обращений к модели	Langfuse, если включена	Задаёт оператор	Кому оператор выдал доступ к Langfuse
Технические журналы	Система сбора журналов в контуре	Задаёт оператор	Оператор
Учётные записи и сессии	Keycloak (своя база PostgreSQL), Redis	Пока учётная запись активна	Администратор Keycloak
Данные источников	Источники заказчика или витрина ClickHouse	Платформа копий не хранит, кроме примеров значений в каталоге и результатов в истории диалогов	Учётная запись платформы только для чтения

Каждая запись и каждый файл принадлежат одной организации — см. «Доступ и аудит» (стр. 19). Шифрования на уровне приложения нет: реквизиты источников и ключи моделей защищены доступом к базе, диски шифрует тот, кто эксплуатирует установку.

Что покидает контур

Только **запросы к языковой модели** — API → шлюз моделей, если он есть → провайдер (см. «Обработка запроса» (стр. 11)), и **интеграции, подключённые явно**, — MCP-серверы, коннекторы, вебхуки (см. «MCP и интеграции» (стр. 17)). Если модель в контуре и интеграций нет, не уходит ничего.

Никогда не покидают контур трассировки, журналы, резервные копии и файлы. Модели и интеграциям не передаются учётные данные источников, токены, идентификаторы и роли пользователей. Вендору не уходит ничего: телеметрии и проверки обновлений нет, лицензия проверяется по подписи в установке.

Персональные данные закрывают скрытием от модели (раздел «Скрытие данных от модели», стр. 15), учётной записью источника с доступом только к нужным колонкам или представлениями без них. Подробности — в приложении «Хранение и удаление данных» (стр. 32).

Скрытие данных от модели

В платформе два потока данных, и идут они разными путями. **Пользователь** получает таблицы, графики и выгрузки: их заполняет сервер прямо из хранилища. **Модель** получает результаты инструментов, из которых собирает ответ, и только этот поток уходит провайдеру. Режим «Скрывать помеченные данные от модели» не пускает помеченные значения во второй поток и не трогает первый.

	Дата	Магазин	Контрагент	Менеджер	Продажи, шт.	Выручка, ₽
Пометка	видно модели	видно модели	только ключи, без названий	скрыто от модели	видно модели	числа метрики скрыты
Видит пользователь всегда, в таблице и выгрузке	02.11.2026	Москва, Тверская 12	7701215046 · ООО «Альфа»	Иванова А. С.	1 240	186 000
Получает модель режим выключен	02.11.2026	Москва, Тверская 12	7701215046 · ООО «Альфа»	Иванова А. С.	1 240	186 000
Получает модель режим «Скрывать помеченные данные» включён	02.11.2026	Москва, Тверская 12	7701215046	`[redacted]`	1 240	`[redacted]`

уходит провайдеру модели заменено перед отправкой в модель

Одна строка результата запроса в трёх представлениях. Поток пользователя режим не меняет.

Как включается

Администратор организации помечает данные в модели данных — у измерения уровни «видно модели», «только ключи», «скрыто», у метрики — видит ли модель её числа — и включает режим в настройках организации. По умолчанию режим выключен; без пометок он ничего не меняет. Пошагово — «Скрытие данных от языковой модели» (см. «Руководство пользователя» → «Руководство администратора» → «Настройки и внешний вид») и «Доступ модели к данным» (см. «Руководство пользователя» → «Руководство администратора» → «Модель данных»).

Что получает модель

Канал	При включённом режиме
Результаты запросов по модели данных	Помеченные значения заменены на `[redacted]`; названия колонок, число строк и сам запрос остаются
Поиск по справочнику	«Только ключи» — ключи без названий; «Скрыто» — поиск недоступен
Произвольный SQL, предпросмотр, схема и статистика источника	Запрещены для всего источника, в котором есть помеченная колонка: SQL пишет модель, и маскировать его по догадке нельзя
Примеры значений колонок в каталоге	Не передаются
Загруженные пользователями файлы	Ассистент их не читает, если в организации что-то помечено
Таблицы, графики, выгрузки пользователя	Не меняются: это поток пользователя

Пометку ассистент ослабить не может: предлагая изменения модели данных, он не может понизить уровень скрытия. Производная метрика из скрытой тоже не может быть видимой.

Ограничения

- Модель не может прокомментировать скрытое число. Ответ — таблица и график, без «выручка выросла на 12 %».
- Режим управляет потоком «хранилище → модель». Если пользователь сам вписал данные в вопрос, они уходят провайдеру.
- При «только ключах» поиск по подстроке работает: модель может подтвердить, что элемент с подходящим ключом существует под названием, которое назвал пользователь.
- Структура модели данных — названия и описания датасетов, метрик, измерений — видна модели всегда. База знаний ограничивается на загрузке документов, не этим режимом.
- Полностью вопрос снимает модель в контуре заказчика — см. «Архитектура» (стр. 5).

MCP и интеграции

Кроме запросов к модели, наружу уходит только то, что передают подключённые явно интеграции. Сразу после установки ни одной интеграции нет. Но функции MCP и вебхуков включены: администратор организации может подключить их сам, без администратора установки. Если это недопустимо, оператор выключает функцию в конфигурации.

Интеграция	Кто подключает	Что уходит наружу	По умолчанию	Как выключить
Свой MCP-сервер организации	Администратор организации, в интерфейсе	Аргументы вызова инструмента, которые выбрала модель; ответ сервера возвращается модели	Функция включена, серверов нет	Вся функция — `api.mcp.enabled: false`; сервер — выключить или удалить в настройках организации
Коннекторы платформы (почта, мессенджер)	Оператор в конфигурации, затем администратор организации включает у себя	Текст сообщения, тема письма, адресаты из разрешённого списка организации	Выключены; в офлайн-дистрибутив не входят	Не включать в чарте и конфигурации или выключить в организации
Вебхуки сигналов	Владелец автопилота или администратор организации	Карточка сигнала: заголовок, текст, разрез, показатели-основания, приоритет, имя автопилота	Функция включена, вебхуков нет	Вся функция — `api.signals.enabled: false`; вебхук — выключить
Внешняя трассировка (LangSmith)	Оператор	Полные трассы: промпты, данные, ответы	Выключена	`api.langsmith.tracing: false`
Продуктовая аналитика (PostHog)	Оператор	События интерфейса с идентификатором пользователя и организации; текст маскируется	Выключена; в дистрибутив не входит	`analytics.enabled: false`

Входящие вебхуки автопилотов только запускают автопилот по токену и наружу ничего не передают.

MCP-серверы

- Только HTTPS (streamable HTTP); локальных процессов платформа не запускает. Новый сервер создаётся выключенным и включается после проверки подключения.

- Заголовки авторизации задаёт администратор; хранятся в базе открытым текстом, через API не возвращаются и в журналы не пишутся.
- Модель сама решает, что передать инструменту, и может отправить фрагменты данных, которые видела в диалоге. Ответ сервера (до 10 000 символов) возвращается модели.
- **Подтверждения пользователя перед вызовом нет**, в том числе для инструментов, которые пишут во внешнюю систему. Подключайте только серверы, которым доверяете это право.
- Не больше 10 серверов и 50 инструментов на организацию, таймаут вызова — 30 секунд.

Ограничения исходящих соединений

Списка разрешённых MCP-серверов на уровне установки нет. Каждое соединение MCP и вебхуков платформа проверяет: только HTTPS; запрещены частные и loopback-адреса и внутренние имена (`.local`, `.svc`, `.cluster.local`); адрес проверяется перед каждым соединением, редиректы не выполняются. Исключения — хосты из списка оператора (`api.mcp.egress.allowHosts`, `api.signals.webhook.egress.allowHosts`); туда же вносят MCP-сервер внутри контура. Конкретные внешние хосты ограничивает межсетевой экран.

Коннекторы почты и мессенджера

Почта — письмо до 10 адресатам только из разрешённого списка организации, не больше 30 в минуту. Мессенджер — сообщение через бота организации в чат из её списка; личные чаты недоступны. Секреты (пароль SMTP, токен бота) модели не показываются; адреса, тексты и токены в журналы не пишутся.

Доступ и аудит

Тема	Как устроено
Вход	Только через Keycloak по OIDC; паролей платформа не хранит. Keycloak установки (можно с LDAP-федерацией или внешним провайдером по OIDC/SAML) или Keycloak заказчика в корпоративном режиме — тогда платформе не нужны права администратора в нём
Точка входа	oauth2-proxy: проверяет сессию до того, как запрос дойдёт до интерфейса или API
Роли	Администратор установки — все организации и панель администратора, назначается в самой платформе; администратор организации (`org-admin`) — участники, источники, модель данных, настройки и история диалогов своей организации; участник (`org-member`) — работа с ассистентом
Организации и роли в корпоративном режиме	Правила сопоставления групп Keycloak заказчика с организацией и ролью; пользователь без подходящей группы не входит
Отзыв доступа	Отключение в Keycloak закрывает вход; исключить участника или заблокировать учётную запись можно и в платформе — это не отменяется следующим входом
Изоляция организаций	Каждая запись и каждый файл принадлежат одной организации. API на каждом запросе сверяет организацию с токеном пользователя; запись чужой организации для запроса не существует. Подключения к источникам и кэш соединений разделены по паре «организация + источник»
Граница изоляции	В общей установке — логическая, в коде приложения, проверяется сквозными тестами тенантности перед каждым выпуском. В выделенной установке от других заказчиков установка отделена физически
Журнал изменений доступа	Организации, правила сопоставления групп, роли, исключения и блокировки: кто, когда, было → стало. Пишется в той же транзакции, что и изменение, приложение записи только добавляет; переживает удаление организации
Другие журналы	Изменения модели данных и схемы источников; история диалогов; трассировки обращений к модели, если включены; технические журналы сервисов
Кто читает диалоги	Администратор своей организации — для разбора качества (участников стоит предупредить); администратор установки — любой
Служебные интерфейсы	Трассировка, журналы, метрики — за единым входом через Keycloak, для отдельной группы администраторов

Подробно — в приложениях «Аутентификация и роли» (стр. 38), «Изоляция тенантов» (стр. 35) и «Журналирование и аудит» (стр. 41).

Сеть и порты

Для on-premise и частного облака: что открыть на межсетевом экране, чтобы установка работала. Соединения между компонентами в кластере правил не требуют. Полный список, включая порты между узлами кластера и доступ на время установки, — в разделе «Требования к инфраструктуре» (см. «Установка и администрирование» → «Требования к площадке»).

Откуда	Куда	Порт	Когда нужно
Пользователи	Вход установки (узел Kubernetes или балансировщик)	443/tcp	Всегда
Интернет	Вход установки	80/tcp	Только для сертификатов Let's Encrypt (проверка HTTP-01)
Пользователи	Keycloak заказчика	443/tcp	Корпоративный вход
Узел и поды платформы	Keycloak заказчика	443/tcp	Корпоративный вход: проверка ключей подписи токенов
Узел и поды платформы	ClickHouse	8123, 9000/tcp	Витрина или источник ClickHouse заказчика
Узел платформы	PostgreSQL заказчика	5432/tcp	Источник PostgreSQL или база платформы у заказчика
Узел платформы	Oracle	1521/tcp	Источник Oracle
Узел платформы	Endpoint языковой модели и модели векторизации	443/tcp	Внешняя модель — по адресу API провайдера (см. приложение «Сетевой периметр» (стр. 43)); своя модель — по её адресу
Узел платформы	Langfuse заказчика	443/tcp	Если трассировка включена
Узел и поды платформы	Машины PostgreSQL, Redis, S3-хранилища установки	5432, 6379, 3900/tcp	Если они вынесены на отдельные машины
Поды платформы	MCP-серверы, получатели вебхуков сигналов	443/tcp	Только если администратор организации их подключил

Исходящий доступ в интернет не нужен, если модели работают в контуре, а сертификаты выпущены заказчиком. Платформа не обращается к серверу вендора. Хосты MCP-серверов и вебхуков платформа не ограничивает — это делает межсетевой экран. Полный перечень исходящих соединений — в приложении «Сетевой периметр» (стр. 43).

Ответы на типовые вопросы ИБ

Ответы ниже относятся к платформе в целом. Где ответ зависит от варианта развёртывания, это сказано; для SaaS конкретные значения — провайдер, сроки, уровень обслуживания — фиксирует договор с оператором.

Какие модели используются — внешние, локальные, смешанные?

Любые из трёх. Платформа подключается к любому провайдеру с OpenAI-совместимым API или другим поддерживаемым API и к собственной модели в контуре заказчика. Основная модель, модель векторизации и распознавание речи настраиваются независимо. В SaaS провайдера выбирает оператор и фиксирует в договоре; в on-premise и частном облаке — заказчик. См. «Модели и маршруты» (стр. 29).

Где размещаются данные?

В SaaS — в установке оператора, в отдельном тенанте. В on-premise — в контуре заказчика; в частном облаке — в выделенной установке. История диалогов, файлы, журналы, трассировки и резервные копии во всех вариантах остаются в установке. См. «Хранение и удаление данных» (стр. 32).

Где данные покидают периметр?

В штатной работе — только в запросах к языковой модели, если модель вне контура, и через инструменты, которые подключил администратор организации (MCP-серверы, коннекторы, вебхуки). Обращений к серверу вендора, телеметрии и проверок обновлений нет. См. «Сетевой периметр» (стр. 43).

Что именно уходит в модель?

Системная инструкция, вопрос с историей диалога, описание схемы источника с примерами значений колонок, результаты SQL-запросов в пределах порога, фрагменты базы знаний. Не уходят идентификаторы, имена и адреса пользователей, роли, учётные данные и токены. См. «Потоки данных» (стр. 24).

Можно ли не передавать модели часть данных?

Да. Администратор организации помечает измерения и метрики в модели данных и включает «Скрывать помеченные данные от модели»: помеченные значения, примеры значений и произвольный SQL по такому источнику модели не передаются, а пользователь видит данные полностью. Режим не скрывает то, что пользователь сам пишет в вопросе. См. «Скрытие данных от модели» (раздел «Потоки данных», стр. 27).

Передаются ли персональные данные?

Платформа рассчитана на агрегированные коммерческие данные. Если источник содержит персональные данные, их закрывают пометками в модели данных, правами учётной записи источника или представлениями без таких колонок. Автоматического распознавания персональных данных нет.

Используются ли данные для обучения моделей?

Платформа не дообучает модели и не содержит такого контура. Использует ли запросы провайдер внешнего API, определяют его условия: в SaaS их проверяет оператор и указывает в договоре, в собственной установке — заказчик при выборе провайдера. С моделью в контуре вопрос не возникает.

Можно ли полностью отказаться от внешних моделей?

Да, в on-premise и частном облаке: все модели, включая векторизацию, размещаются в контуре, и платформа работает без доступа в интернет. Качество ответов зависит от выбранной модели и оборудования — это обсуждается до установки. См. «Изолированный контур» (раздел «Сетевой периметр», стр. 44).

Как разделены данные разных заказчиков?

Каждая запись принадлежит одной организации, и принадлежность проверяется на каждом запросе по серверному контексту, а не по параметрам клиента; чужие записи для запроса не существуют. Это проверяет отдельный набор сквозных тестов. Если логической изоляции недостаточно, выбирают выделенную установку. См. «Изоляция тенантов» (стр. 35).

Как устроен вход и можно ли использовать корпоративные учётные записи?

Вход — через Keycloak по OIDC: либо Keycloak установки с федерацией корпоративного каталога или внешним провайдером, либо Keycloak заказчика в корпоративном режиме, где платформе не нужны административные права. См. «Аутентификация и роли» (стр. 38).

Что журналируется и может ли заказчик это проверить?

Изменения доступа (организации, роли, членство, блокировки) — с автором, временем и содержанием изменения; изменения модели данных; история диалогов. Администратор организации читает диалоги своей организации, администратор установки — журнал изменений доступа. См. «Журналирование и аудит» (стр. 41).

Шифруются ли данные?

Каналы — по HTTPS. Диски и тома шифруются на уровне инфраструктуры, это настраивает оператор. Дополнительного шифрования на уровне приложения нет. См. «Шифрование» (раздел «Хранение и удаление данных», стр. 33).

Что происходит при недоступности модели?

Сетевые ошибки повторяются до двух раз; автоматического переключения на другую модель нет. Резервную модель администратор установки включает вручную — её стоит завести и проверить заранее.

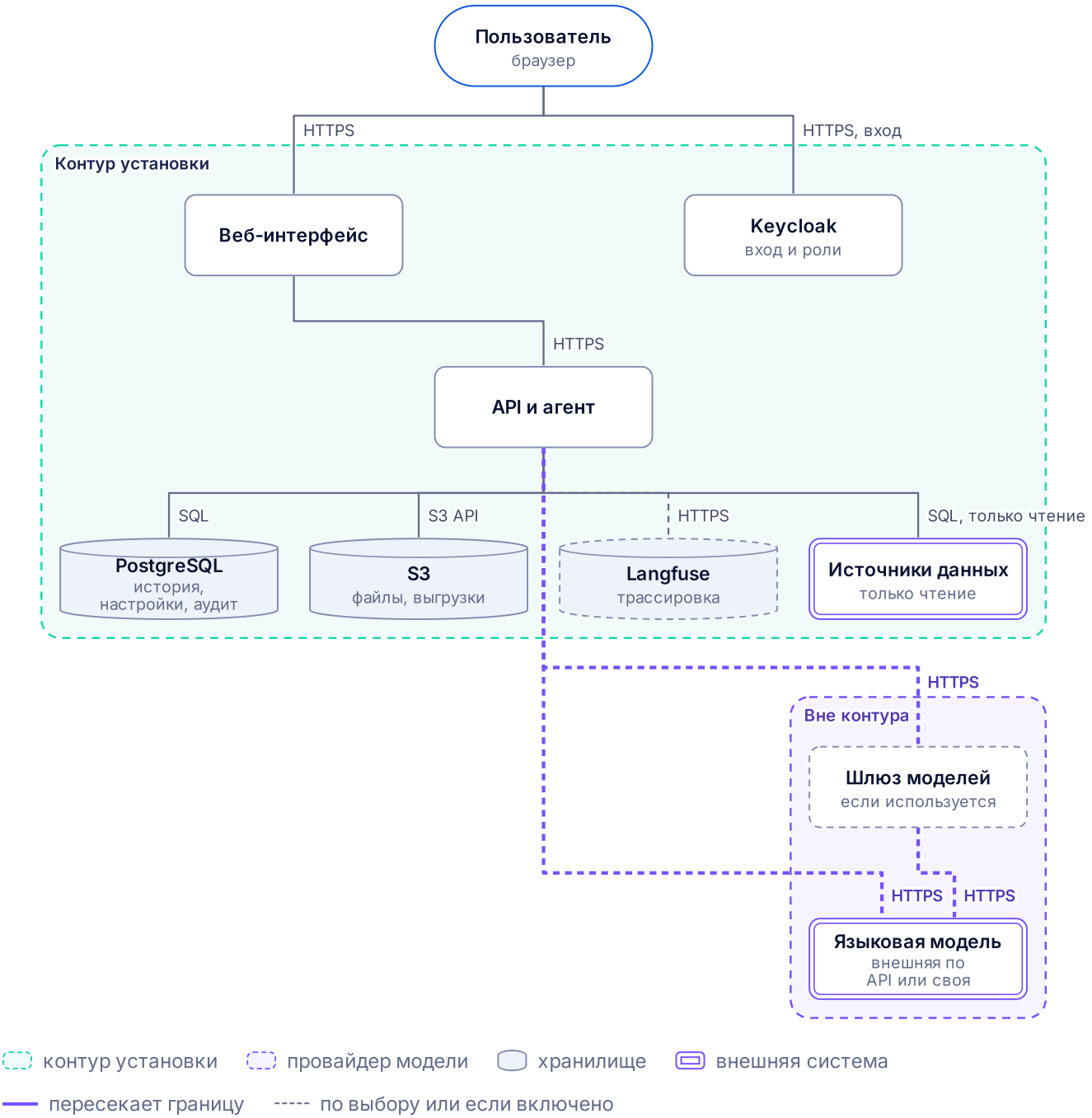
Как удаляются данные?

Диалоги удаляет пользователь, файлы к вопросам удаляются через 30 дней, выгрузки — через сутки. Удаление всех данных организации выполняет оператор; в резервных копиях данные остаются до истечения их срока. См. «Хранение и удаление данных» (стр. 32).

Потоки данных

Участники обмена

Компонент	Где размещён	Роль
Веб-интерфейс и API	Контур установки	Вход, проверка прав, работа ассистента, хранение истории
PostgreSQL	Контур установки	История диалогов, настройки организаций, база знаний и её векторный индекс, журналы изменений
Объектное хранилище (S3-совместимое)	Контур установки	Файлы пользователей, крупные результаты запросов, выгрузки
Источники данных заказчика: ClickHouse, PostgreSQL, Oracle	Контур установки или инфраструктура заказчика	Данные для аналитики; доступ по учётной записи, которую выдал заказчик
Трассировка обращений к модели (Langfuse)	Контур установки; необязательна	Разбор ответов ассистента и контроль качества
Шлюз моделей	По выбору оператора; необязателен	Маршрутизация запросов к провайдеру, учёт и лимиты расходов
Провайдер модели	Внешний API или модель в контуре заказчика	Ответы ассистента, векторизация текстов



Платформа, её хранилища и источники данных работают в контуре установки; наружу уходят только запросы к языковой модели.

Как обрабатывается вопрос

Один вопрос пользователя обычно приводит к нескольким обращениям к модели: план ответа, SQL-запрос, разбор результата.

1. Пользователь задаёт вопрос в веб-интерфейсе.
2. API передаёт модели системную инструкцию, вопрос с историей диалога и описание доступных данных.
3. Модель возвращает план и SQL-запрос.

4. API выполняет запрос в источнике данных учётной записью только для чтения.
5. Результат возвращается модели: целиком, если он небольшой, или предпросмотр, если он крупный, — тогда полный результат сохраняется в объектном хранилище внутри контура.
6. Модель формулирует ответ; API сохраняет его в историю диалога и показывает пользователю.

Что передаётся в языковую модель

Категория	Содержание	Откуда
Системная инструкция	Правила работы ассистента; описание организации, если его задал администратор организации; каталог модели данных — названия и описания датасетов, метрик, измерений	Настройки организации
Вопрос пользователя	Текст вопроса и история текущего диалога	Пользователь
Описание источника	Названия таблиц и колонок, типы, описания и **примеры значений колонок** — до 50 значений на колонку, сохранённые в каталоге при описании источника	Каталог схемы
Результаты SQL-запросов	Строки результата, если их меньше порога; выше порога — предпросмотр и сводка, полный результат остаётся в объектном хранилище	Источник данных
Фрагменты базы знаний	Части документов организации, найденные по вопросу	База знаний организации
Сохранённые запросы	SQL-запросы, которые пользователи ранее подтвердили как верные	PostgreSQL платформы

Порог задаётся для каждой модели в панели администратора (см. «Установка и администрирование» → «Панель администратора установки»): по умолчанию результат крупный, если в нём от 100 строк или от 100 КБ, а предпросмотр — 20 строк.

Не передаётся в модель: идентификаторы пользователя и организации, имя, адрес электронной почты, роли, учётные данные источников, токены доступа. Они хранятся в серверном контексте запроса, и модель не может ни прочитать их, ни подставить в параметры своих инструментов.

Другие обращения к провайдеру

Функция	Что передаётся	Когда
Векторизация текстов	Тексты документов базы знаний и описаний схемы при индексировании, текст вопроса при поиске	При загрузке документов, описании источника, поиске
Распознавание речи	Аудиозапись голосового ввода	Только если функция включена в конфигурации и пользователь говорит, а не пишет

Функция	Что передаётся	Когда
Описание источника	Те же сведения, что в таблице выше, и примеры значений, которые ассистент читает из источника, чтобы описать колонки	При подключении источника и обновлении описания

Модели для векторизации и распознавания речи настраиваются отдельно от основной и могут работать у другого провайдера или в контуре — см. «Модели и маршруты» (стр. 29).

Скрытие данных от модели

Помеченные значения можно не пускать в модель, оставив их пользователю, — см. главу «Скрытие данных от модели» (стр. 15).

Персональные данные

Платформа рассчитана на агрегированные коммерческие данные: продажи, остатки, товары, точки продаж, контрагенты. Если в источнике есть персональные данные, они могут попасть в модель через примеры значений колонок и через результаты запросов. Автоматического распознавания персональных данных в платформе нет. Меры:

1. пометить такие колонки в модели данных и включить скрытие данных от модели (стр. 15);
2. выдать платформе учётную запись источника с доступом только к нужным таблицам и колонкам;
3. подключать представления (view) без персональных данных вместо исходных таблиц;
4. исключить такие таблицы из описания источника при подключении.

Где данные хранятся внутри контура

Подробно, со сроками по умолчанию — на странице «Хранение и удаление данных» (стр. 32). Коротко: история диалогов, настройки и база знаний — в PostgreSQL; файлы и крупные результаты — в объектном хранилище с префиксом организации и пользователя в ключе; трассировки — в Langfuse внутри контура, если трассировка включена.

Точки выхода данных

В штатной работе содержимое диалогов, фрагменты документов и данные источников покидают контур установки только одним путём: **API → шлюз моделей, если он есть → провайдер модели.** Если модель размещена в контуре, этот путь тоже остаётся внутри. Остальные исходящие

соединения — все необязательные и выключенные по умолчанию — перечислены на странице «Сетевой периметр» (стр. 43).

Модели и маршруты

Какие модели использует платформа

Назначение	Где настраивается	Обязательна
Основная модель: ответы ассистента, SQL, разбор результатов	Реестр моделей в панели администратора (см. «Установка и администрирование» → «Панель администратора установки»), раздел «Модели LLM»	Да
Векторизация текстов для поиска по базе знаний и описаниям схемы	Конфигурация `api`, параметр `embeddings`	Да
Распознавание голосового ввода	Конфигурация `api`, параметр `stt`	Нет: без него голосовой ввод выключен

Каждая из них может работать у своего провайдера. Например, основная модель — через внешний API, а векторизация — моделью в контуре.

Поддерживаемые провайдеры

Способ подключения	Что подключается
OpenAI-совместимый API	Любой провайдер, шлюз моделей или агрегатор с OpenAI-совместимым API, а также сервер модели в контуре заказчика
Нативный API провайдера	Коммерческие провайдеры, у которых платформа поддерживает их собственный API; перечень — по запросу

Модель в контуре заказчика подключается через OpenAI-совместимый сервер или через необязательный сервис `kafka-llm-proxy`, если доступ к моделям в контуре организован через Kafka. Для векторизации дополнительно поддерживаются модели, развёрнутые в контуре.

Реестр моделей

Перечень моделей, доступных ассистенту, ведёт администратор установки в панели администратора. Для каждой модели задаются провайдер, идентификатор модели, эндпоинт, ключ доступа и параметры генерации, а также пороги объёма данных, передаваемых модели:

Параметр	По умолчанию	Что значит
Порог по строкам	100	Результат запроса от этого числа строк считается крупным
Порог по объёму	100 КБ	То же по объёму

Параметр	По умолчанию	Что значит
Строк в предпросмотре	20	Сколько строк крупного результата получает модель

Крупный результат целиком сохраняется в объектном хранилище внутри контура, модель получает только предпросмотр и сводку. Чем ниже пороги, тем меньше данных уходит провайдеру.

Перед сохранением модель можно проверить кнопкой проверки подключения — она отправляет минимальный тестовый запрос. Ключи доступа после сохранения в интерфейсе не показываются.

Администратор организации выбирает модель по умолчанию для своей организации только из этого перечня — см. «Настройки» («Руководство пользователя» → «Руководство администратора» → «Настройки и внешний вид»).

Отказ модели

Автоматического переключения на другую модель в платформе нет. При сетевой ошибке запрос к той же модели повторяется до двух раз с нарастающей паузой; ошибки, которые вернул сам провайдер (например, превышение лимита), не повторяются этим механизмом. Если модель недоступна, пользователь получает сообщение об ошибке.

Переключение на резервную модель — ручное действие администратора установки: сменить модель в реестре или модель по умолчанию организации. Чтобы оно было быстрым, резервную модель стоит завести в реестре заранее и проверить.

Лимиты расходов

Расход на модели можно ограничить бюджетом организации или отдельных участников — см. «Бюджет» («Руководство пользователя» → «Руководство администратора» → «Бюджет»). Если перед провайдером стоит шлюз моделей, он может дополнительно задавать отдельный ключ и лимиты на каждую организацию; это настройка шлюза, а не платформы.

Выбор провайдера по вариантам развёртывания

	SaaS	On-premise и частное облако
Кто выбирает провайдера	Оператор; перечень моделей и маршрутов фиксируется в договоре	Заказчик

	SaaS	On-premise и частное облако
Можно ли отказаться от внешних моделей	Нет: SaaS работает с провайдером, указанным в договоре	Да: все модели, включая векторизацию, можно разместить в контуре
Условия провайдера о хранении и обучении на данных	Оператор проверяет и указывает в договоре	Заказчик проверяет по договору с выбранным провайдером
Изменение перечня	По согласованию с заказчиком	Решение заказчика

Платформа не дообучает модели на данных заказчика: в ней нет такого контура. Использует ли провайдер запросы для обучения, определяют условия провайдера — их стоит проверить до подключения внешнего API.

Хранение и удаление данных

Где что хранится и сколько

Сроки в таблице — поведение платформы по умолчанию. Там, где срок задаёт не платформа, а оператор установки, это указано.

Данные	Хранилище	Срок по умолчанию
История диалогов: вопросы, ответы, SQL, результаты в объёме, который видела модель	PostgreSQL	Пока диалог не удалит пользователь
Рабочее состояние ассистента между шагами диалога	PostgreSQL	Для каждого диалога хранятся два последних состояния; у диалога без активности 30 дней состояние удаляется целиком. История при этом не затрагивается
Настройки организации, модель данных, подключения к источникам	PostgreSQL	Пока их не удалит администратор
База знаний и её векторный индекс	PostgreSQL	Пока документы не удалят
Файлы, прикреплённые к вопросам	Объектное хранилище	30 дней, затем автоматическое удаление
Выгрузки результатов в файл	Объектное хранилище	24 часа; ссылка на скачивание действует 15 минут
Крупные результаты запросов, файлы библиотеки	Объектное хранилище	Срока по умолчанию нет; оператор задаёт его политикой жизненного цикла хранилища
Трассировки обращений к модели	Langfuse в контуре	Задаёт оператор в Langfuse; если трассировка выключена — не создаются
Журнал изменений доступа	PostgreSQL	Бессрочно, переживает удаление организации
Технические журналы сервисов	Система сбора журналов в контуре	Задаёт оператор
Поток ответа ассистента	NATS JetStream	Минуты
Кэш и сессии	Redis	До перезапуска

Файлы, прикреплённые к вопросам, удаляются в два шага: по истечении срока они перестают быть доступны, а из хранилища стираются через сутки.

> ⚠ **Осторожно**
>
> Трассировки содержат промпты, ответы модели и фрагменты данных из источников. Обращайтесь с ними как с пользовательскими данными: ограничьте доступ к Langfuse и задайте срок хранения.

Резервные копии

Резервное копирование — ответственность оператора установки. Обязательно копировать PostgreSQL (основную базу, базу Keycloak и базу трассировки) и объектное хранилище; рабочее состояние ассистента можно не копировать — без него дамп восстанавливается корректно.

- Если PostgreSQL ставится инсталлятором на отдельную виртуальную машину, в комплект входит скрипт резервного копирования, который запускается по таймеру.
- Для PostgreSQL внутри кластера резервное копирование не настраивается: нужны снимки тома или `pg_dump` по расписанию.

Резервная копия содержит данные всех организаций установки. Порядок, состав и проверка восстановления — на странице «Данные, бэкапы и retention» (см. «Установка и администрирование» → «Данные, бэкапы и retention»).

Удаление

- **Диалог** удаляет пользователь в интерфейсе.
- **Учётная запись**: отключение или удаление пользователя в Keycloak закрывает вход, но не удаляет историю его диалогов — её удаляют отдельно, см. «Пользователи, роли и доступ» («Установка и администрирование» → «Пользователи, роли и доступ»).
- **Источник данных**: отключение источника удаляет всё, что на нём построено, — см. «Отключить источник» («Руководство пользователя» → «Руководство администратора» → «Подключение источников данных»).
- **Массовое удаление** — например, всех данных организации по окончании договора — выполняет оператор на стороне базы данных и хранилища.

После удаления из основной базы данные остаются в резервных копиях до истечения их срока хранения.

Шифрование

Где	Как
Между браузером и платформой	HTTPS; TLS терминируется на балансировщике перед кластером
Между платформой и провайдером модели	HTTPS
Между платформой и источниками данных	Средствами СУБД; сертификат корпоративного удостоверяющего центра для Oracle и других источников указывается при установке — см. «Доменные имена и TLS» («Установка и администрирование» → «Требования к площадке»)

Где	Как
Данные на дисках	Шифрование дисков и томов на уровне инфраструктуры — настраивает оператор
На уровне приложения	Не применяется: реквизиты подключений и ключи моделей хранятся в базе платформы и защищены доступом к ней

Изоляция тенантов

Модель изоляции

Одна установка обслуживает несколько организаций — тенантов. Сервисы и хранилища у них общие, но каждая запись принадлежит ровно одной организации, и принадлежность проверяется на каждом запросе на уровне приложения.

Где проходит граница, зависит от варианта развёртывания:

Уровень	Тенант в общей установке (SaaS)	Выделенная установка (on-premise, частное облако)
Сервисы платформы	Общие с другими заказчиками	Только заказчика
PostgreSQL	Общая база, организация указана в каждой записи	Своя база
Объектное хранилище	Общий бакет, организация в начале ключа объекта	Свой бакет
Источники данных	Подключения привязаны к организации	То же
Keycloak	Общий, организация = Organization в Keycloak	Свой Keycloak или Keycloak заказчика
Резервные копии, журналы, трассировки	Общие, с идентификатором организации в записях	Свои

В выделенной установке изоляция тенантов разделяет подразделения или юрлица одного заказчика; от других заказчиков установка отделена физически.

Как проверяется принадлежность

1. Пользователь входит через Keycloak и получает токен со списком своих организаций.
2. На каждом запросе API проверяет, что запрошенная организация есть в этом списке. Если нет — запрос отклоняется. Заголовок с «самообъявленной» организацией ничего не даёт.
3. Идентификаторы пользователя, организации и роли записываются в серверный контекст запроса. Клиент и языковая модель изменить их не могут.
4. Любое чтение или изменение записи сверяет её организацию с организацией контекста. Запись чужой организации для запроса не существует: ответ «не найдено», без признака того, что такая запись есть.

Исключение — администратор установки: он может войти в любую организацию. Это право назначают только тем, кто эксплуатирует установку, см. «Аутентификация и роли» (стр. 39).

Политики безопасности на уровне строк СУБД для таблиц платформы не используются: граница проходит в коде приложения и проверяется сквозными тестами (ниже).

Источники данных

Подключение к источнику хранится как пара «организация + источник» со своими учётными данными. Соединение ищется по обоим ключам, и кэш соединений тоже разделён по этой паре, поэтому учётные данные другой организации даже не загружаются в обработку запроса.

Некоторые источники подключает администратор установки — например, витрину, общую для нескольких организаций. Какой организации доступен какой источник, задаёт привязка в панели администратора. Реквизиты таких подключений администратор организации не видит.

Для аналитических хранилищ в контуре установки рекомендуется дополнительная граница на стороне СУБД: отдельная база и отдельный пользователь только для чтения на каждую организацию. Платформа их не создаёт — это настройка СУБД, которую выполняет оператор.

База знаний и файлы

- Документы базы знаний и их векторный индекс хранятся в PostgreSQL с идентификатором организации. Поиск идёт только в пределах организации из серверного контекста, а не из параметров запроса.
- Ключ объекта в хранилище файлов начинается с идентификаторов организации и пользователя. При каждом обращении к файлу платформа дополнительно сверяет организацию в его метаданных с организацией сессии.

Пользователи и роли

Пользователь может состоять в нескольких организациях и в каждой иметь свою роль: администратор одной и участник другой. Администратор организации управляет участниками, источниками и настройками только своей организации. Подробно — на странице «Аутентификация и роли» (стр. 38).

Как проверяется отсутствие межклиентского доступа

1. **Сквозные тесты тенантности** — отдельный набор, который проверяет: отказ в доступе к организации, где пользователь не состоит; игнорирование самообъявленной организации в заголовке; отказ администратору организации в чтении и изменении данных другой

организации, в том числе той, где он состоит, но в которую не вошёл; отказ обычному участнику в административных функциях.

2. **Проверка выпуска** — набор прогоняется перед выпуском версии.
3. **Журнал изменений доступа** — создание и изменение организаций, ролей и членства записываются, см. «Журналирование и аудит» (стр. 41).
4. **Проверка заказчиком** — заказчик может провести собственное тестирование; для этого нужны две тестовые организации.

Ограничения

- Изоляция в общей установке логическая: общая СУБД, общий бакет, общие сервисы. Ошибка в коде приложения теоретически может открыть доступ к чужим данным; это компенсируется тестами и журналами. Если это неприемлемо, выбирают выделенную установку.
- Дополнительного шифрования на уровне приложения нет. Реквизиты подключений к источникам и ключи доступа к моделям хранятся в базе платформы и защищены доступом к ней и шифрованием дисков — его включает оператор на уровне инфраструктуры.

Аутентификация и роли

Учётные записи и пароли платформа не хранит: вход идёт через Keycloak по OIDC. Платформа поддерживает два режима, и выбор между ними — главное решение о входе.

Режимы входа

	Keycloak установки	Корпоративный режим
Чей Keycloak	Разворачивается вместе с платформой	Заказчика
Где заводятся пользователи	В Keycloak установки; можно подключить корпоративный каталог	В системах заказчика
Откуда берутся организации и роли	Организации и группы Keycloak установки	Правила сопоставления групп заказчика в платформе
Какие права нужны платформе в Keycloak	Администратор Keycloak установки	Никаких административных: только клиент для входа
Подходит, если	Нет требований к единому корпоративному входу или достаточно федерации каталога	Учётными записями должна управлять только служба заказчика

Keycloak установки

Каждая организация — Organization в Keycloak. Роли в организации задаются членством в её группах. Как заводить пользователей и выдавать роли — на странице «Пользователи, роли и доступ» (см. «Установка и администрирование» → «Пользователи, роли и доступ»).

Корпоративные учётные записи подключаются средствами Keycloak:

- федерация каталога по LDAP или LDAPS: пользователи живут в корпоративном каталоге, а членство в организациях задаётся сопоставлением групп каталога с группами организаций;
- вход через внешний провайдер идентификации по OIDC или SAML.

Корпоративный режим

Платформа подключается к Keycloak заказчика как обычное приложение и читает из токена перечень групп пользователя. Организацию и роль по группам определяют правила сопоставления, которые администратор установки ведёт в панели администратора («Правила доступа»):

- правило связывает путь группы или шаблон пути с организацией и ролью;
- пользователь, ни одна группа которого не подошла ни к одному правилу, в платформу не попадает;

- первый администратор установки задаётся в конфигурации — группой, членство в которой даёт это право;
- проверить правила до их сохранения можно симуляцией: она показывает, что получит пользователь с заданным набором групп.

От заказчика нужны конфиденциальный клиент OIDC и перечень групп с полными путями в токене. Платформа хранит у себя только тех пользователей, которые хотя бы раз вошли.

Роли

Роль	Что даёт
Администратор установки	Доступ ко всем организациям и к панели администратора установки. Список администраторов ведётся в самой платформе, а не ролью в Keycloak (см. ниже). Назначается только тем, кто эксплуатирует установку
Администратор организации (`org-admin`)	Участники, источники данных, модель данных, настройки и история диалогов своей организации
Участник (`org-member`)	Работа с ассистентом, дашборды, библиотека и файлы в своей организации

Администраторы установки хранятся в базе платформы и назначаются в панели администратора теми, у кого уже есть это право. Первого администратора задаёт конфигурация: в корпоративном режиме — группа Keycloak заказчика, членство в которой даёт это право (её можно убрать из конфигурации, когда администраторы назначены в платформе). В установках, где администраторов раньше определяла роль `admin` в Keycloak установки, она по-прежнему учитывается как один из источников этого списка, но сама по себе прав не даёт.

Роли организации действуют в контексте выбранной организации: один и тот же пользователь может быть администратором одной организации и участником другой. Что видит и делает администратор организации, описано в руководстве администратора (см. «Руководство пользователя» → «Руководство администратора» → «Администрирование организации»).

Сессии и отзыв доступа

- Отключение пользователя в Keycloak закрывает вход; в корпоративном режиме изменения групп в Keycloak заказчика действуют со следующего входа или обновления сессии.
- Исключить участника из организации или заблокировать учётную запись можно и в самой платформе — это действует и в корпоративном режиме и не отменяется следующим входом.
- Каждое такое изменение записывается в журнал изменений доступа (раздел «Журналирование и аудит», стр. 41).

- Перед служебными веб-интерфейсами — трассировкой, журналами, Grafana — стоит единая точка входа через Keycloak; доступ к ним выдаётся отдельной группе администраторов.

Журналирование и аудит

Какие журналы есть

Журнал	Что содержит	Где смотреть	Кто видит
Журнал изменений доступа	Создание, изменение, включение и удаление организаций; правила сопоставления групп; выдача и снятие ролей; исключение участника из организации и блокировка учётной записи. По каждой записи — кто, когда и что именно изменилось (было → стало)	Панель администратора → «Правила доступа» → вкладка «Аудит»	Администратор установки; записи фильтруются по правам смотрящего
Журнал изменений модели данных	Публикация, правка и отклонение датасетов, метрик и измерений, с причиной отклонения	Модель данных организации	Администратор организации
Журнал изменений схемы источника	Добавленные и удалённые таблицы и колонки	Карточка источника → «Таблицы»	Администратор организации
История диалогов	Вопросы, ответы, оценки и комментарии участников	«История диалогов» — см. «Контроль качества» («Руководство пользователя» → «Руководство администратора» → «Контроль качества»)	Администратор организации — по своей организации; администратор установки
Трассировки обращений к модели	Промпты, ответы, вызовы инструментов, длительность и стоимость; идентификаторы пользователя, диалога и организации	Langfuse в контуре	Те, кому оператор выдал доступ к Langfuse
Технические журналы сервисов	События сервисов, ошибки	Система сбора журналов в контуре	Оператор

Журналы изменений доступа и модели данных приложение только дополняет — оно не изменяет и не удаляет записи. Запись в журнал изменений доступа делается в той же транзакции, что и само изменение, поэтому изменения без записи не бывает. Записи остаются и после удаления организации.

В корпоративном режиме входа, когда пользователи и группы живут в Keycloak заказчика, журнал изменений доступа — единственная запись о том, кто и когда получил доступ к какой организации (см. «Аутентификация и роли» (стр. 38)).

Кто может читать диалоги

Администратор организации читает все диалоги участников своей организации — это нужно для разбора жалоб на качество ответов. Участников стоит об этом предупредить. Администратор установки читает диалоги любой организации.

Доступ к журналам и трассировкам

Трассировки и история диалогов содержат тексты вопросов и фрагменты данных. Рекомендуется:

- выдавать доступ к Langfuse, системе сбора журналов и Grafana отдельной группе администраторов, а не участникам организаций;
- задать срок хранения трассировок и технических журналов, см. «Хранение и удаление данных» (стр. 32);
- если трассировка не нужна для разбора качества, держать её выключенной — по умолчанию она выключена.

Выгрузка для аудита

Выгрузку журнала изменений доступа, истории диалогов и трассировок за период выполняет оператор установки: в on-premise — ИТ заказчика, в SaaS — оператор по запросу заказчика.

Сетевой периметр

Входящие соединения

Пользователи и администраторы обращаются к установке только по HTTPS через балансировщик, на котором терминируется TLS:

- веб-интерфейс и API — включая WebSocket для потоковой выдачи ответа;
- Keycloak — страница входа;
- служебные интерфейсы — трассировка, журналы, метрики, оркестратор пайплайнов, консоль хранилища. Их стоит публиковать только во внутреннюю сеть администраторов.

Сервисы платформы и хранилища общаются между собой внутри кластера и снаружи не публикуются. Требования к балансировщику и доменным именам — в разделе «Требования к инфраструктуре» (см. «Установка и администрирование» → «Требования к площадке»).

Исходящие соединения

Куда	Что уходит	Нужно
Провайдер языковой модели или шлюз моделей: адреса API провайдера выдаются при подключении	Содержимое запросов к модели — см. «Потоки данных» (стр. 24)	Да, если модель вне контура
Провайдер векторизации	Тексты документов, описаний схемы и вопросов	Да, если модель векторизации вне контура
Сервис распознавания речи	Аудио голосового ввода	Только если голосовой ввод включён
Источники данных заказчика	SQL-запросы; обратно — результаты	Да, если источник вне кластера
Корпоративный каталог или Keycloak заказчика	Вход пользователей	Только при корпоративном входе
MCP-серверы и встроенные коннекторы (мессенджер, почта)	Параметры вызова инструмента или текст сообщения — то, что ассистент решил передать	Только если их подключил администратор организации; коннекторы выключены по умолчанию
Вебхуки сигналов	Карточка сигнала: заголовок, текст, разрез, показатели	Только если они настроены
Внешний сервис трассировки	Промпты, ответы, идентификаторы	Выключено по умолчанию; трассировка в контуре (Langfuse) не требует выхода наружу
Сервис продуктовой аналитики	События интерфейса с именем пользователя и организацией; введённый текст маскируется	Выключено по умолчанию

Платформа **не** обращается к серверу вендора: лицензия проверяется в установке по подписи, без связи с внешним сервером; проверок обновлений и отправки телеметрии вендору нет. При установке и обновлении офлайн-дистрибутивом доступ к внешнему реестру образов тоже не нужен.

Точки выхода данных

Содержимое диалогов, фрагменты документов и данные источников в штатной работе покидают контур только в двух случаях:

1. **запрос к языковой модели** — путь API → шлюз моделей, если он есть → провайдер модели. Если все модели размещены в контуре, этого выхода нет;
2. **инструменты, которые подключил администратор организации** — MCP-серверы, коннекторы и вебхуки. Они передают наружу то, для чего их подключили.

Трассировка, журналы, резервные копии и файлы остаются внутри контура.

Изолированный контур

Платформа работает без доступа в интернет, если:

- основная модель и модель векторизации размещены в контуре и доступны по OpenAI-совместимому API — см. «Модели и маршруты» (стр. 29);
- распознавание речи выключено или тоже работает в контуре;
- внешняя трассировка и продуктовая аналитика не включены;
- MCP-серверы и коннекторы, если нужны, работают внутри контура; их адреса оператор вносит в `egress.allowHosts`, иначе платформа не подключится к внутренним адресам — см. «MCP и интеграции» (стр. 17).

Тогда на периметре достаточно разрешить входящий HTTPS к балансировщику и соединения с источниками данных.

Условия для модели SaaS

SaaS: обработка данных и SLA

> **Условия модели SaaS**
>
> Эта страница описывает условия, на которых DeepDive предоставляет DeepDive как сервис. В on-premise и частном облаке они не действуют: там конфигурацию, сроки и обслуживание определяет заказчик и его договор с вендором. Окончательные условия фиксирует договор; при расхождении действует договор.

Конфигурация установки

Параметр	Значение
Размещение установки	Дата-центр Selectel, Российская Федерация
Языковые модели	Один провайдер коммерческих моделей, платный корпоративный API: ответы ассистента, векторизация, распознавание речи. Другие провайдеры не подключены
Шлюз моделей	Инфраструктура оператора, вне контура заказчика: тела запросов не сохраняет; ключ и лимит расходов на организацию
Трассировка	Langfuse в установке, 30 дней
Технические журналы	90 дней
Резервные копии PostgreSQL	Ежедневно, хранятся 14 дней; проверочное восстановление — еженедельно
Файлы пользователей и крупные результаты запросов	30 дней
Аналитические хранилища в установке	Отдельная база и пользователь только для чтения на каждую организацию, ограничительные политики строк

Путь данных при обращении к модели: установка (Selectel, Россия) → шлюз моделей → провайдер модели. Это единственная точка выхода содержимого диалогов из установки. Юрисдикции шлюза и провайдера — в таблице субпроцессоров ниже.

Модели и маршруты

Точный идентификатор основной модели фиксируется в приложении к договору на дату подключения и меняется только по согласованию.

Маршрут	Когда	Как
Основной: основная модель провайдера через шлюз моделей	Штатно	—
Резервный 1: резервная модель того же провайдера через шлюз моделей	Недоступность или деградация основной модели	Переключает инженер DeepDive; если модель отличается от зафиксированной — по согласованию с заказчиком
Резервный 2: API провайдера напрямую, в обход шлюза	Недоступность шлюза	Переключает инженер DeepDive

Автоматического переключения нет; нормативы ручного переключения — в разделе «Отказ провайдера моделей» ниже. Если доступ к API провайдера ограничит сам провайдер или регулятор, DeepDive уведомляет заказчика в течение 1 часа и предлагает альтернативу; данные иному провайдеру передаются только после письменного согласия заказчика.

По условиям платного корпоративного API провайдера запросы и ответы не используются для обучения и улучшения моделей. DeepDive проверяет актуальность условий при каждом выпуске пакета документов.

Другие провайдеры подключаются только по письменному согласованию с заказчиком; собственная модель в контуре доступна в on-premise как отдельный проект.

Обработка данных

Роли сторон

DeepDive обрабатывает данные заказчика только для оказания услуги по договору и по его поручению. Заказчик определяет цели и состав данных, в том числе какие источники подключить.

Категории данных

Категория	Примеры	Цель
Данные пользователей	Имя, рабочий адрес электронной почты, роль, идентификатор в системе входа заказчика	Вход, проверка прав, журнал действий
Содержимое диалогов	Вопросы, ответы, SQL-запросы, результаты в объёме, переданном модели	Оказание услуги, история, контроль качества
Метаданные источников	Названия таблиц и колонок, описания, примеры значений	Модель данных и построение запросов
Данные источников	Коммерческие данные, к которым заказчик дал доступ	Аналитические запросы пользователей
Документы базы знаний	Регламенты, справочники, описания	Поиск и ответы с учётом документов
Файлы пользователей	Загрузки и выгрузки	Оказание услуги
Технические данные	Журналы, метрики, трассировки с идентификаторами пользователя и организации	Эксплуатация, диагностика, безопасность
Данные для расчётов	Число пользователей, объём использования моделей	Биллинг

Персональные данные покупателей заказчика по умолчанию не обрабатываются. Если источники их содержат, стороны фиксируют это при подключении и применяют меры со страницы «Потоки данных» (стр. 27).

Субпроцессоры

Субпроцессор	Назначение	Какие данные получает	Юрисдикция
Провайдер коммерческой языковой модели	Языковые модели: ответы, векторизация, распознавание речи	Содержимое запросов к модели, тексты для векторизации, аудио голосового ввода	По договору
Хостинг установки	Размещение установки: вычисления, диски, объектное хранилище, резервные копии	Все данные на уровне инфраструктуры, без доступа к содержимому	Россия
Хостинг шлюза моделей	Размещение шлюза моделей	Запросы и ответы модели в транзите, без хранения; учёт по ключам	По договору

Перечень субпроцессоров с наименованиями — приложение к договору, предоставляется по запросу.

Субпроцессорами **не** являются: сервисы аналитики поведения пользователей, внешние системы поддержки с доступом к данным, внешние службы резервного копирования, внешние сервисы трассировки.

Изменение субпроцессоров и провайдеров моделей

1. DeepDive уведомляет заказчика о добавлении или замене субпроцессора не менее чем за 30 календарных дней: назначение, состав данных, страна, условия.
2. Заказчик вправе направить мотивированное возражение в течение 14 календарных дней. Стороны ищут альтернативу — иной провайдер, выделенная установка, отключение функции; если её нет, заказчик вправе расторгнуть договор в части затронутой услуги без штрафных санкций.
3. Экстренная замена при недоступности провайдера — только из согласованных резервных маршрутов; иные замены — после уведомления и согласия заказчика.

Сроки хранения и удаления

Данные	В период действия договора	После прекращения
История диалогов	Пока её не удалит пользователь или администратор организации	30 дней на выгрузку по запросу, затем удаление в течение 30 дней
База знаний	Период действия договора	Удаление в течение 30 дней
Файлы и крупные результаты запросов	30 дней с момента создания	По истечении срока
Метаданные источников и реквизиты подключений	Период действия договора	Реквизиты — в день прекращения доступа, метаданные — в течение 30 дней

Данные	В период действия договора	После прекращения
Учётные записи пользователей	Период действия договора	Отключение в день прекращения, удаление в течение 30 дней
Журнал изменений доступа	Период действия договора	3 года в сокращённом виде: действие, дата, тип объекта, без содержимого
Технические журналы	90 дней	Истекают по сроку
Трассировки	30 дней	Истекают по сроку
Резервные копии	14 дней	Данные исчезают из копий не позднее 14 дней после удаления из основной базы

Факт удаления DeepDive подтверждает письменно по запросу: дата и перечень удалённых категорий.

Инциденты безопасности

- Инцидент — подтверждённый или обоснованно предполагаемый несанкционированный доступ к данным заказчика, их утрата, изменение или раскрытие, включая инциденты у субпроцессоров.
- Уведомление — не позднее 24 часов с момента подтверждения, контактным лицам из договора: описание, затронутые данные, принятые меры, контакт DeepDive.
- Итоговый отчёт — в течение 10 рабочих дней: причины, объём, меры по устранению и предотвращению.
- DeepDive содействует заказчику в уведомлении регуляторов и субъектов данных.

Аудит

- Заказчик администрирует свою организацию: участники, настройки, история диалогов. Выгрузку журнала изменений доступа за период DeepDive предоставляет по запросу.
- Не чаще раза в год и дополнительно после инцидента DeepDive по запросу предоставляет: конфигурацию маршрутов моделей без секретов, протокол проверки того, что шлюз не сохраняет тела запросов, протокол последней проверки резервного маршрута, изменения в перечне субпроцессоров за период.
- Заказчик вправе провести аудит сам или силами независимого аудитора по согласованной программе, с уведомлением за 30 дней, в рабочее время, с соблюдением конфиденциальности других заказчиков.
- Доступ заказчика не распространяется на данные других заказчиков, внутренние журналы инфраструктуры и системные журналы субпроцессоров.

Обязательства сторон

DeepDive: обрабатывать данные только по поручению и в целях договора; не использовать их для обучения моделей без письменного согласия; обеспечивать меры безопасности из этого раздела; привлекать только согласованных субпроцессоров; уведомлять об инцидентах; удалять данные по окончании договора; содействовать аудиту.

Заказчик: определять состав источников и не давать доступ к данным, обработка которых запрещена законом или его договорами; выдавать учётные записи источников с минимальными правами; управлять участниками своей организации; сообщать об изменении контактных лиц.

Уровень обслуживания

Доступность

Целевая доступность — **99,5 % за календарный месяц**, круглосуточно: не более 3 часов 36 минут простоя в 30-дневном месяце.

Платформа доступна, если пользователь может войти, отправить вопрос и получить ответ. Это проверяет внешний мониторинг DeepDive контрольным сценарием не реже раза в минуту из точки вне инфраструктуры платформы. Простой — минуты, когда сценарий завершался ошибкой не менее чем в двух проверках подряд:

> Доступность = (минуты месяца − минуты простоя) / минуты месяца × 100 %

В простой не входят:

1. плановые работы в согласованном окне с уведомлением не менее чем за 2 рабочих дня;
2. недоступность по причинам на стороне заказчика: его источники данных, сеть, учётные записи, система входа;
3. недоступность провайдера моделей, если резервный маршрут включён в пределах норматива;
4. обстоятельства непреодолимой силы;
5. ограничение доступа к провайдеру моделей по решению государственных органов или операторов связи.

Плановые работы

- Окно — воскресенье, 03:00–06:00 по московскому времени.
- Уведомление — не менее чем за 2 рабочих дня; срочные работы по безопасности — не менее чем за 4 часа.
- Не более 8 часов в месяц. Обновления без прерывания сервиса окном не ограничены.

Отказ провайдера моделей

Ситуация	Действие DeepDive	Норматив
Основная модель отвечает ошибкой или деградировала	Переключение на резервный маршрут	Начало работ в пределах реакции P1; переключение — в течение 4 часов в часы поддержки
Все согласованные маршруты недоступны	Уведомление заказчика, предложение альтернативы	Уведомление — в течение 1 часа
Провайдер изменил условия обработки данных	Уведомление до переключения трафика; переключение только после согласования	—

Поддержка

Каналы — электронная почта поддержки и выделенный чат для контактных лиц заказчика; для P1 — телефон дежурного инженера. Адреса и телефоны указываются в договоре. Стандартные часы — рабочие дни, 09:00–19:00 по московскому времени; P1 — круглосуточно.

Приоритет	Определение	Реакция	Решение или обходное решение
P1 — критический	Платформа или организация недоступна для всех пользователей; утечка или подозрение на утечку	30 минут, круглосуточно	4 часа
P2 — высокий	Недоступна ключевая функция: вопросы ассистенту, подключение источника; существенная деградация	2 часа	1 рабочий день
P3 — средний	Ошибка у части пользователей или сценариев, есть обходное решение	1 рабочий день	5 рабочих дней
P4 — низкий	Вопрос, пожелание, косметический дефект	2 рабочих дня	По плану выпусков

Реакция — время от регистрации обращения до подтверждения, что оно принято и ему присвоен приоритет. Для P2–P4 считается в часы поддержки.

Компенсации

Если доступность за месяц ниже целевой, заказчик вправе получить скидку на абонентскую плату следующего месяца:

Фактическая доступность	Скидка от месячной платы
от 99,0 % до 99,5 %	5 %
от 98,0 % до 99,0 %	10 %
ниже 98,0 %	25 %

Запрос направляется в течение 30 дней после окончания месяца; DeepDive в течение 10 рабочих дней предоставляет отчёт мониторинга и подтверждает или мотивированно отклоняет запрос. Компенсации за месяц не превышают 25 % месячной платы и являются единственным средством защиты заказчика в части доступности.

Отчётность и пересмотр

По запросу DeepDive ежемесячно предоставляет отчёт: фактическая доступность, инциденты с длительностью и причинами, плановые работы, переключения на резервные маршруты моделей. Условия пересматриваются по инициативе любой стороны не чаще раза в 6 месяцев, если иное не вызвано изменением состава провайдеров моделей.

Сборка 8817d38 от 3 ноября 2026 г.