Cognitum AI Platform · администратору

0.17.0 → **0.17.1** 23 ноября 2026

Вход через свой IdP, файлы до 4 ГБ

Для администратора: корпоративный вход с группами из каталога, файлы в обе стороны с MCP, каталог MCP и модели таблицами, расход API по людям и настраиваемый пул соединений с базой.



Вход и учёт
SSO, расход API, пул БД

Скиллы
Категории и заявки

Файлы
До 4 ГБ, файлы MCP

Инструменты
Каталог MCP, модели

4
образа обновить: приложение, агенты, исполнитель

3
уровня подтверждения вызовов MCP

0
правок в агентах на SDK 0.17.0

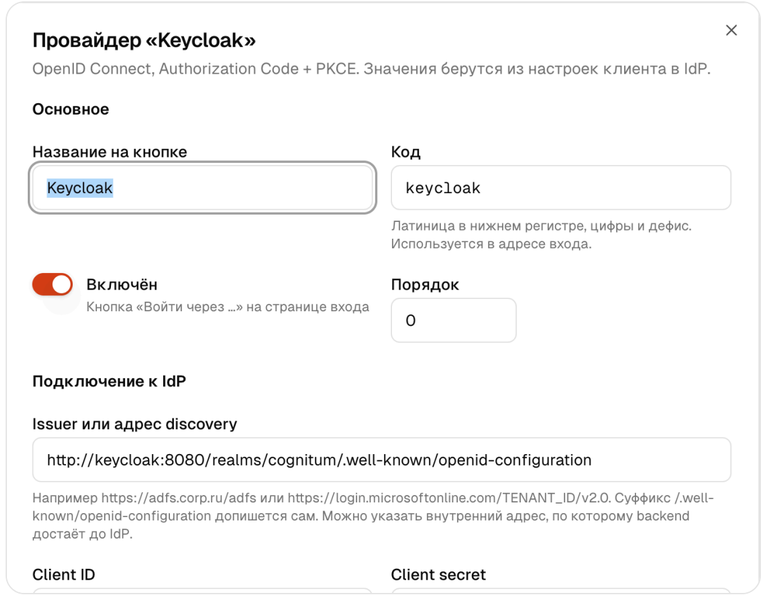
SSO настраивается в одном диалоге

ПРОБЛЕМА

Сотрудники входили паролем платформы или через LDAP. Контроллеры домена на новых Windows Server закрывают внешний LDAP, поэтому нужен был вход по OpenID Connect.

ЧТО ИЗМЕНИЛОСЬ

На вкладке «SSO» в «Администрировании» подключается провайдер по OpenID Connect: ADFS (Windows Server 2019 и новее), Microsoft Entra ID или Keycloak. Форма логина и пароля остаётся аварийным входом, если провайдер недоступен.



Диалог провайдера: название кнопки, код и адрес провайдера.

ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Проверьте, что `PUBLIC_BASE_URL` — https-адрес платформы: из него собирается адрес возврата.
2. Зарегистрируйте клиента в IdP, добавьте провайдера на вкладке «SSO» и нажмите «Проверить».
3. Проверьте вход в приватном окне, не закрывая окно администратора.
4. Сохраните `APP_SECRET_KEY` в резервной копии: им зашифрован секрет клиента.

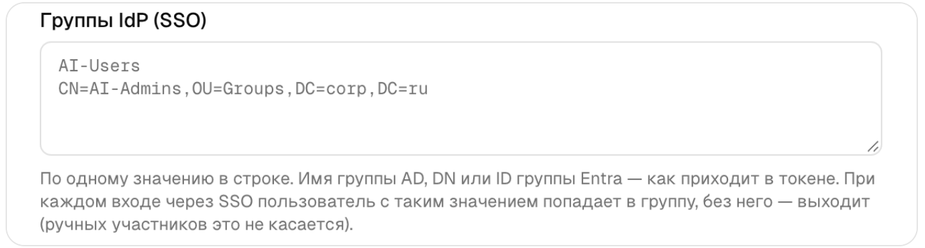
Группы и роли приходят из каталога

ПРОБЛЕМА

Группы и роли новым сотрудникам администратор раздавал вручную — и так же вручную снимал, когда человек уходил из отдела.

ЧТО ИЗМЕНИЛОСЬ

У группы и роли есть поле «Группы IdP (SSO)». При каждом входе платформа сверяет его с тем, что прислал провайдер: исключили из группы в каталоге — роль снимется при следующем входе. Ручные назначения и назначения из LDAP не трогаются.



Поле «Группы IdP (SSO)» у группы платформы.

ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Укажите, где провайдер передаёт группы: `groups` у Keycloak и Entra, `roles` у ролей приложения Entra, `group` у ADFS.
2. Впишите в группы и роли платформы значения «Группы IdP (SSO)» — так, как их присылает провайдер.
3. «Пускать только членов сопоставленных групп» включайте после сопоставления — остальные увидят «Нет доступа к платформе».
4. Роль не пришла — откройте карточку пользователя: «Группы из IdP при последнем входе» покажут, что прислал провайдер.

Скриншот со стенда, тестовая группа

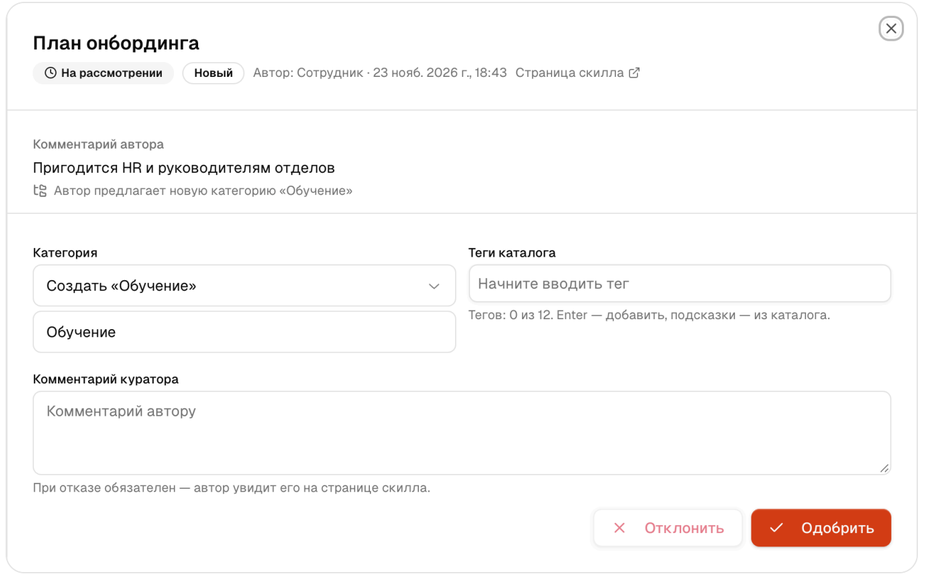
Куратор разбирает заявки в одной панели

ПРОБЛЕМА

Каталог рос одной лентой для всех отделов. Куратор узнавал о новой заявке, только зайдя на страницу, а обновление опубликованного скилла приходилось сравнивать с текущей версией вручную.

ЧТО ИЗМЕНИЛОСЬ

Каталог раскладывается по категориям с доступом — тем же диалогом «Доступ», что у моделей. Заявка открывается в боковой панели: текст, файлы, а для обновления — сравнение «сейчас / предлагается». О новых заявках приходит уведомление.



Заявка в боковой панели: автор, категория и решение куратора.

ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Заведите категории в «Каталоге скиллов» на вкладке «Категории» и ограничьте доступ там, где полка нужна не всем.
2. Разбирайте заявки в боковой панели: отказ — только с причиной, автор увидит её на странице скилла.
3. Обновление скилла принимайте по сравнению «сейчас / предлагается» — автор увидит решение у себя.

Скриншот со стенда, тестовая заявка

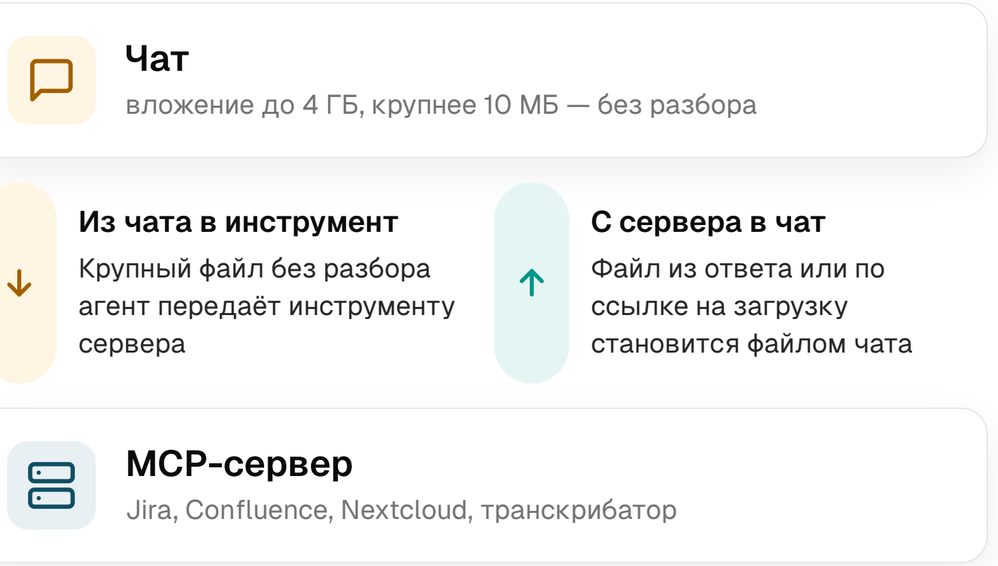
Файлы до 4 ГБ — и в обе стороны с MCP

ПРОБЛЕМА

Частями грузились только аудио и видео, остальное упиралось в лимит 50 МБ, и каждое вложение сразу уходило в разбор, даже если агенту нужно было только передать файл в инструмент. Внешний MCP-сервер мог вернуть файл только внутри ответа.

ЧТО ИЗМЕНИЛОСЬ

Любой файл больше обычного лимита грузится частями до 4 ГБ, а вложение чата крупнее 10 МБ сохраняется без разбора. Файл из ответа MCP-сервера становится файлом чата, а сервер может сам залить файл в платформу по ссылке на загрузку.



ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. `FILES_LARGE_MAX_UPLOAD_MB` задаёт потолок, по умолчанию 4096 МБ. Старое имя `FILES_MEDIA_MAX_UPLOAD_MB` ещё читается.
2. Заложите место в резервном копировании: том файлов растёт — файлы до 4 ГБ любого типа.
3. Для Jira в режиме «комментарии и вложения» обновите MCP-сервер Atlassian до 0.23.1-ite.4.
4. Разработчику MCP-сервера: пометьте аргумент файла `x-cognitum-upload` — ссылку подставит платформа.

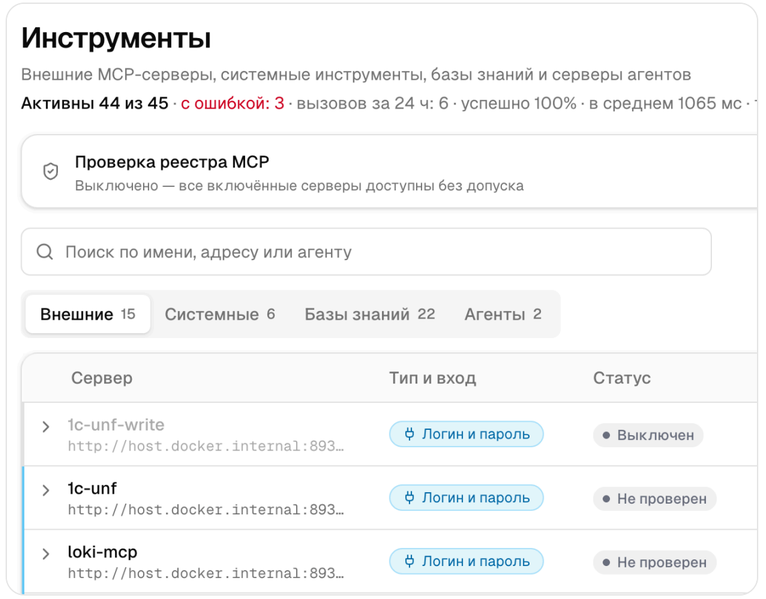
Каталог MCP читается как таблица моделей

ПРОБЛЕМА

Каталог MCP в администрировании был лентой карточек. Чтобы найти сервер с ошибкой или сравнить серверы, приходилось открывать каждую карточку.

ЧТО ИЗМЕНИЛОСЬ

«Инструменты» в администрировании устроены как «LLM API»: сводка за 24 часа, поиск, вкладки «Внешние», «Системные», «Базы знаний», «Агенты» и таблица. После трёх отказов входа подряд подключение само уходит в офлайн.



Сводка за 24 часа, поиск и вкладки каталога MCP.

ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Обновите MCP-сервер Nextcloud до 0.3.1 — он сообщает платформе об отказе входа.
2. Офлайн-подключение возвращается новыми учётными данными или ручной проверкой.
3. Проблемные серверы видны в сводке «с ошибкой» — не нужно открывать каждую карточку.

Скриншот со стенда, адреса серверов тестовые

Подтверждение вызовов — в три уровня

ПРОБЛЕМА

Подтверждение настраивал только администратор. Но у сервиса с личным входом список инструментов появляется после входа сотрудника — администратор его не видит, и настроить подтверждение было некому.

ЧТО ИЗМЕНИЛОСЬ

Сотрудник в карточке своего подключения может добавить подтверждение. В чате «Разрешить в этом чате» снимает его до конца беседы. Инструменты только для чтения подтверждения не спрашивают.

Администратор · MCP-сервер

Задаёт минимум: «никогда», «для изменяющих» или «всегда».

сотрудник может только добавить

Сотрудник · своё подключение

Добавляет подтверждение для подключения и для отдельных инструментов.

только в живом разговоре

Чат · «Разрешить в этом чате»

Снимает подтверждение до конца беседы, в том числе требование администратора.

ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Решите, где подтверждение обязательно, — в своих настройках сотрудник его не снимет.
2. «Разрешить в этом чате» не действует на автоматизации, фоновые запуски и «Ждут моего решения».
3. Выдача и отзыв разрешений пишутся в аудит — видно, кто и что разрешил.

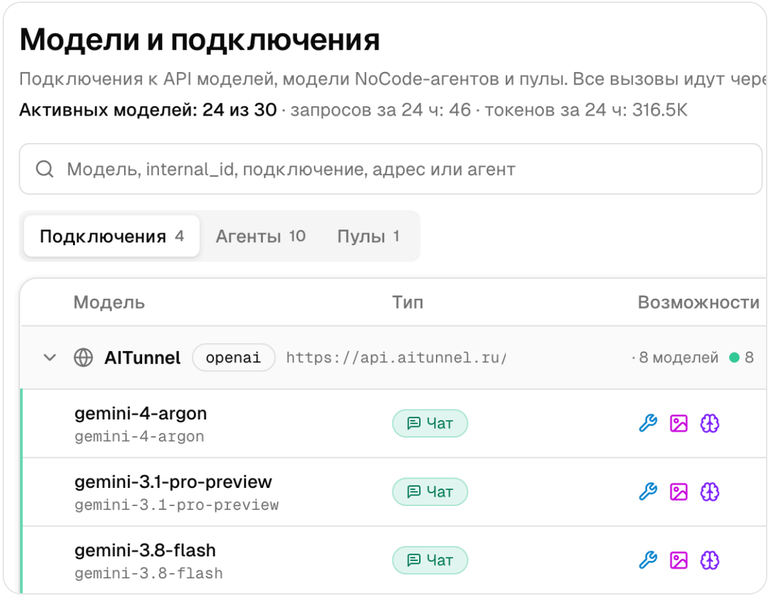
«Модели и подключения» — вкладки, поиск, таблица

ПРОБЛЕМА

На установке с сотней NoCode-агентов «LLM API» превращалась в длинную ленту без поиска и сравнения. Чтобы найти модель конкретного агента, приходилось долго листать.

ЧТО ИЗМЕНИЛОСЬ

Страница разложена по вкладкам «Подключения», «Агенты» и «Пулы» с поиском и компактной таблицей. «Продолжить» в пуле моделей дописывает ответ на модели, где хватит места.



Вкладка «Подключения»: поиск и модели одной таблицей.

ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Ничего настраивать не нужно: «Продолжить» в пуле работает само — в чате и у NoCode-агентов.
2. Вкладка и поисковый запрос сохраняются в адресе страницы — ссылкой можно поделиться с коллегой.
3. Пункт меню прежний — «LLM API», а заголовок страницы теперь «Модели и подключения».

Скриншот со стенда, адреса и ключи скрыты

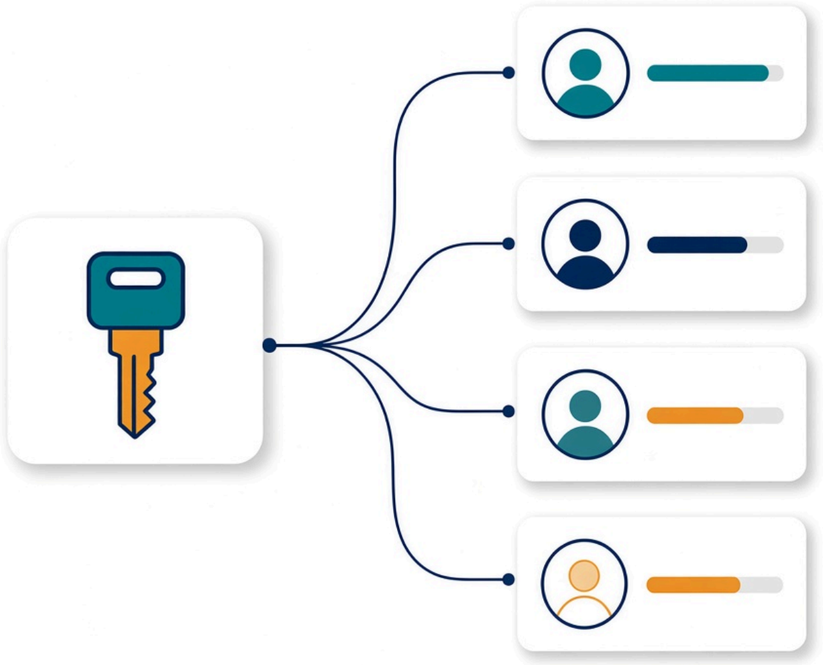
Один токен интеграции — расход по людям

ПРОБЛЕМА

Внешнюю систему вроде Open WebUI или CRM подключают одним токеном, а работают в ней разные люди. Весь её расход в «Потреблении» лежал на владельце токена.

ЧТО ИЗМЕНИЛОСЬ

Интеграция передаёт с запросом, от чьего имени вызов. Если почта совпала с сотрудником, вызов записывается на него, иначе — на внешнего пользователя отдельной строкой. Доступ к модели, тариф и плательщик по-прежнему определяет владелец токена.



ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Включите у токена интеграции признак «Может указывать конечного пользователя» — в «Моих API-ключах» или «Токенах доступа».
2. Передавайте `X-Cognitum-User-Email` и `X-Cognitum-User-Name` или поле `user` — для чата, эмбеддингов, переранжирования и речи.
3. Токенам Open WebUI с вызовами за последние 90 дней признак уже включён при обновлении.

Настраиваемый пул и самопроверка соединений

ПУЛ СОЕДИНЕНИЙ

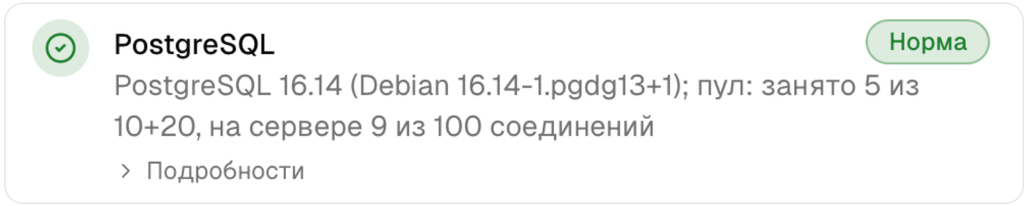
Размер пула соединений приложения с PostgreSQL теперь задаётся в `.env` — переменными `DB_POOL_SIZE` и `DB_MAX_OVERFLOW`, по умолчанию 10 + 20. Его можно подобрать под лимит соединений сервера и число агентов.

САМОПРОВЕРКА

В «Самопроверку» добавлена проверка по количеству соединений: строка PostgreSQL показывает, сколько соединений занято в пуле приложения и сколько — на сервере.

Лимит соединений сервера PostgreSQL





Строка PostgreSQL в «Самопроверке»: сколько занято в пуле и на сервере.

ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Пересоздайте `postgres` вместе с остальными контейнерами — у сервера новые параметры.
2. Пул вместе с соединениями агентов и запасом в 5 должен быть меньше `POSTGRES_MAX_CONNECTIONS`.
3. Внешнему PostgreSQL задайте таймаут зависших транзакций 30 минут и keepalive вручную.

Скриншот со стенда, задержка скрыта

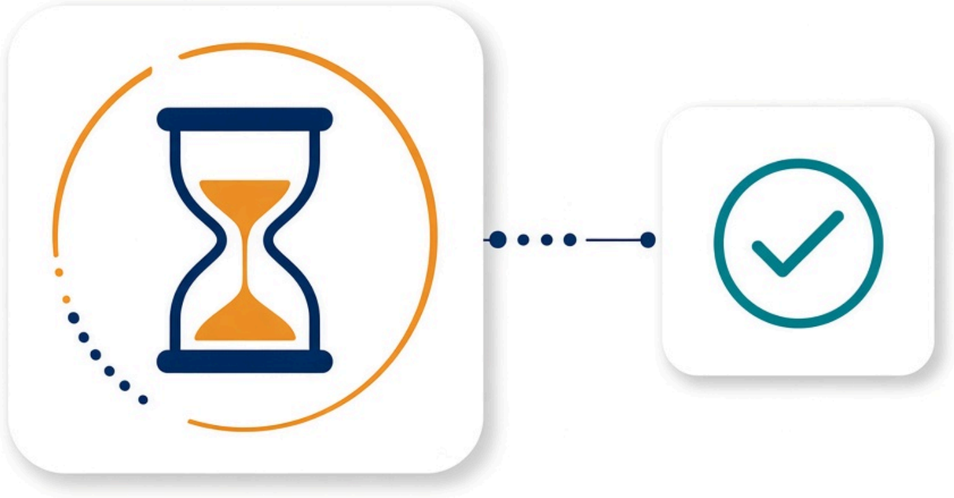
Агент ждёт задачу, виджет виден в просмотре

ПРОБЛЕМА

Поставив долгую задачу MCP-серверу, модель опрашивала статус подряд или заканчивала ход словами «проверьте позже». Показать поддержке виджет MCP-приложения в просмотре чата было нельзя.

ЧТО ИЗМЕНИЛОСЬ

Модель ждёт в том же ходе до 2 минут, а ожидание задачи MCP-сервера заканчивается, как только она завершилась. В просмотре чата виджет открывается в режиме «Только просмотр» или заменяется заглушкой с причиной.



ЧТО СДЕЛАТЬ ПОСЛЕ ОБНОВЛЕНИЯ

1. Ожидание работает само — в чате Workspace и у NoCode-агентов, когда в ходе есть MCP-инструменты.
2. Чтобы виджет показывался в просмотре, обновите сервер до SDK 0.17.1 и поддержите контракт `cognitum/view`.
3. Пометьте инструменты чтения `readOnlyHint`: только они доступны в просмотре и не требуют подтверждения.

Как обновиться: четыре образа и два файла

1. Возьмите из дистрибутива 0.17.1 `docker-compose.yml` и `nats-entrypoint.sh`.
2. Обновите образы `app`, `nocode-harness`, `echo-agent` и `tool-runner` до 0.17.1.
3. Пересоздайте контейнеры **вместе с** `postgres` **и** `nats` — миграции применятся при старте.

```
docker compose pull && docker compose up -d --force-recreate
```

4. Внешнему PostgreSQL задайте таймаут зависших транзакций и keepalive — команды в инструкции обновления.
5. Внешние MCP-серверы — по желанию: Atlassian 0.23.1-ite.4, Nextcloud 0.3.1, транскрибатор 0.1.3, экспортёр метрик 0.9.12, Loki 0.1.0.

Резервное копирование

- Том файлов растёт: теперь в нём файлы до 4 ГБ любого типа.
- Секрет клиента SSO зашифрован `APP_SECRET_KEY` — без ключа вход через SSO не восстановить.

Совместимость

Агенты на SDK 0.17.0 работают без правок. Нужны PostgreSQL 12+, NATS с JetStream, Docker 24+ и Compose V2.

Исправлено 13 досадных ошибок — в управлении соединениями с базой данных, в работе с общими токенами и в других местах

Поправлены потенциальные ошибки в части прав/полномочий или информационной безопасности