Nevel

БЕЗОПАСНОСТЬ

Эксплуатация, резервирование и восстановление

Профили доступности, RPO и RTO, резервные копии, аварийное восстановление, обновления и проверяемая приёмка.

Облако / В контуре

https://docs.nevel.sh/security/operations/

Эксплуатационный профиль Nevel определяет, какие компоненты обслуживаются, от каких отказов защищена установка, сколько данных допустимо потерять и за какое время восстанавливается работа. Профиль связывает договорные показатели с архитектурой и испытаниями. Наличие резервной копии, реплики или автоматического переключения по отдельности не доказывает достижение всех показателей.

В облаке инфраструктуру своей поставки обслуживает Nevel. В on-prem заказчик обеспечивает ресурсы установленной платформы, хранилища, сеть и эксплуатационный доступ в соответствии с распределением работ. Локальные модели обслуживаются внутри его инфраструктуры; российские и глобальные провайдеры и self-hosted модели Nevel имеют отдельную область эксплуатации и доступности. Запуск на одном сервере и многоплощадочная установка имеют разные области отказа. Выбор варианта описан в методах развёртывания.

Какие показатели входят в профиль

Базовый профиль Guardrail версии 1.2 содержит следующие значения. Они описывают предусмотренный профиль сервиса; применимость к конкретному окружению и составу поставки устанавливается его паспортом и договором. Таблица не объявляет эти значения измеренным результатом всех установок Platform или безусловным обязательством каждого тарифа.

Показатель	Значение профиля	Что входит в измерение
Доступность	99,9%	Доступность определённого сервиса за установленный период
Задержка API p50	Не более 150 мс	Обработка API без времени ответа LLM
Задержка API p99	Не более 500 мс	Та же граница, 99-й процентиль
Плановое окно недоступности	Не более 30 минут, уведомление за 48 часов	Окно обслуживания по применимому порядку
Реакция на критический инцидент	Не более 15 минут	Начало реакции по установленному каналу и критичности
RTO	Не более 1 часа	Время до восстановления предусмотренной работоспособности
RPO	Не более 5 минут	Максимальный разрыв между событием отказа и восстановленной точкой данных

Для доступности нужны точное определение успешной операции, интервал наблюдения и источник измерения. В паспорте указываются затронутые компоненты, часовой пояс, правила учёта частичной недоступности и плановых работ. Исключения внешнего провайдера или обслуживания не возникают автоматически: они применяются только в установленной области договора.

Численно 99,9% за условный год из 365 суток соответствует 8,76 часа недоступности. Это арифметическое пояснение, а не разрешение суммировать любой простой за любой период. Годовая доля и ежемесячное обязательство рассчитываются по разным периодам. Реакция за 15 минут также не означает устранение причины за 15 минут.

Задержка API без LLM отделяется от времени до первого токена и полного ответа модели. Большой файл, длинный вывод и очередь модели влияют на пользовательское ожидание по-разному. Порог DLP `p95 > 500 мс` из мониторинга относится к другому этапу и процентилю; его нельзя использовать как доказательство соблюдения `p99` всего API.

Резервирование и разрешённое переключение

Облачный профиль предусматривает несколько российских площадок, переключение при отказе, горизонтальное масштабирование API Gateway, DLP и Router, а также репликацию хранилищ соответствий и аудита. Переключение модели выполняется только на разрешённого получателя. В on-prem внешний резерв допустим только при заранее разрешённом для данных маршруте и с записью решения в аудит. В изолированном профиле или для метки «только в контуре» отказ локальной модели не разрешает отправку во внешний API.

Установка проверяется по фактическим областям отказа. Два процесса на одном сервере не защищают от отказа сервера; две машины на одном хранилище не защищают от отказа этого хранилища. Отдельно рассматриваются питание, сеть, зона размещения, управляющие учётные записи и система ключей. Архитектурные границы и обработка отказа защиты описаны в архитектуре Guardrail.

Отказ	Ожидаемое эксплуатационное действие	Что проверить
Экземпляр Gateway или DLP	Исключить неисправный экземпляр и использовать здоровый	Защита не пропущена, повтор не раскрывает данные
Локальная модель	Использовать разрешённую альтернативу либо вернуть контролируемый отказ	Нет внешнего переключения для запрещённой категории
Основное хранилище	Перейти по предусмотренному плану восстановления	Подтверждена актуальность данных и единственный активный источник записи
Площадка	Выполнить процедуру аварийного переключения	Доступны зависимости, ключи, маршруты и аудит
Логическое повреждение	Восстановить состояние до повреждения	Ошибка не воспроизведена из реплики или копии

Наличие настроенного переключения проверяется отказом соответствующего компонента на стенде или в разрешённом испытании. Отключение только пользовательского интерфейса не подтверждает переживание отказа базы данных.

Копии, репликация и RPO

Профиль резервного копирования Guardrail предусматривает инкрементальные копии каждые 4 часа, полные ежедневно, хранение 30 дней, отдельную российскую площадку, AES-256 и ежемесячное восстановление с протоколом. Состав включает конфигурацию, политики, аудит и действующие соответствия. Подробные условия хранения и удаления приведены в жизненном цикле данных; набор резервируемых объектов Platform определяется составом поставки.

Реплика помогает пережить физический отказ, но может немедленно повторить ошибочное удаление или повреждение. Резервная копия сохраняет отдельное состояние, однако интервал 4 часа сам по себе не обеспечивает RPO 5 минут. Для этого показателя необходим подтверждённый механизм более свежего восстановления: например, предусмотренная архитектурой репликация либо восстановление по журналу изменений с достаточной частотой и глубиной хранения. Его наличие и результат проверяются, а не выводятся из расписания копий.

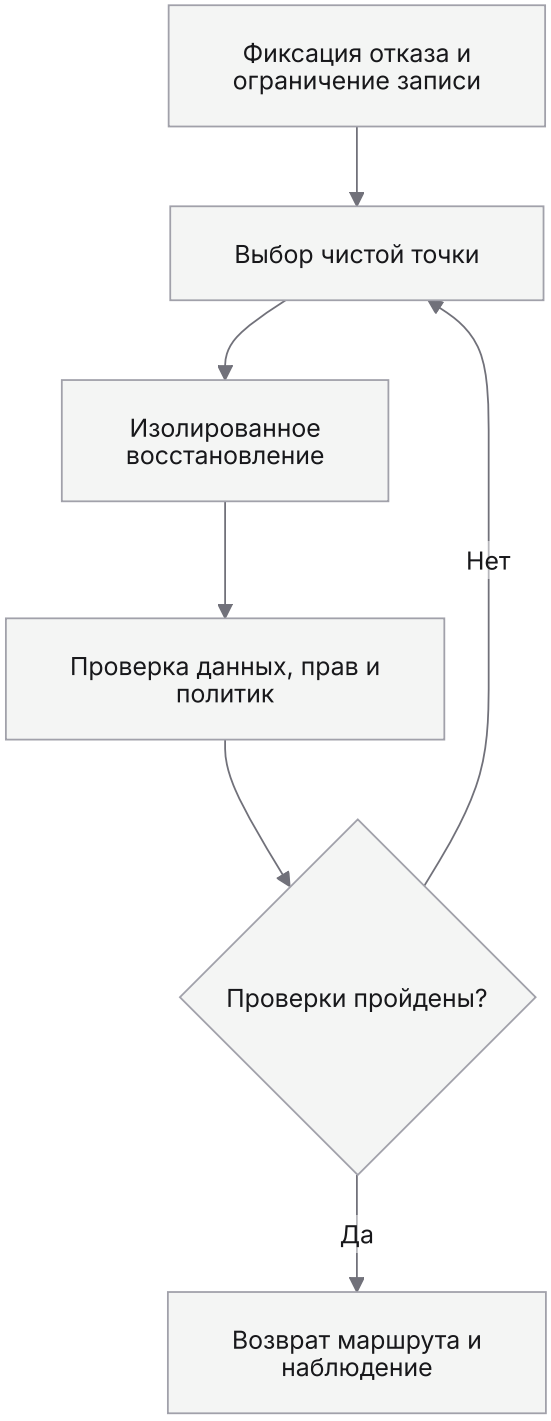
В учебном примере сервис отказал в 14:20. Последняя подтверждённая запись в восстановленном состоянии относится к 14:17, работа возвращена в 15:05. Потеря по времени — 3 минуты, восстановление — 45 минут. Эти значения укладываются в указанный профиль для данного испытания. Если пригодна только копия от 12:00, разрыв составляет 2 часа 20 минут: успешный запуск из неё не подтверждает RPO 5 минут.

Копия проверяется на полноту, читаемость и пригодность ключа расшифрования. Ключи и права восстановления не должны зависеть от единственного отказавшего сервиса. При этом хранение ключа рядом с общедоступным архивом уничтожает смысл шифрования. Требования к ключам и криптографическим границам находятся в разделе о шифровании.

Процедура аварийного восстановления

СХЕМА 1

Аварийное восстановление Nevel



1. Зафиксировать начало воздействия, симптомы, последнее достоверное состояние и решение о восстановлении. При признаках атаки выполнять параллельно процедуру инцидента.
2. Ограничить запись и фоновые задачи там, где они могут создать конкурирующие состояния. Определить единственный источник, который будет принят за рабочий после восстановления.
3. Выбрать точку с учётом причины отказа. Для логического повреждения последняя по времени копия может уже содержать ошибку.

4. Восстановить в изолированном окружении конфигурацию и необходимые хранилища. Не включать внешние отправки и автоматические интеграции до проверки маршрутов.
5. Проверить согласованность объектов: файл доступен разрешённому пользователю, индекс соответствует правам, журнал связан с операциями, маски не восстанавливаются по чужим соответствиям.
6. Применить предусмотренный контроль удалений и отзывов, чтобы копия не вернула удалённые данные, старые права или скомпрометированный ключ.
7. Выполнить положительные и отрицательные тесты доступа и Guardrail, затем вернуть рабочий маршрут. Проверить накопленные задачи и риск повторной отправки.
8. Измерить фактические RPO и RTO, описать потери и ограничения, сохранить протокол и продолжить усиленное наблюдение.

Откат инфраструктуры не отменяет уже состоявшийся внешний вызов. Повтор задачи может создать второй документ, платный запрос или внешнее действие. Поэтому восстановление очередей требует решения, какие операции безопасно повторять, какие сопоставлять по идентификатору, а какие разбирать вручную.

Обновления и изменения конфигурации

Перед обновлением составляется карточка изменения: компоненты и версии, миграции, затронутые политики, зависимости, ожидаемый эффект, проверка успеха и условия отката. Для on-prem проверяются пакеты, образы, контроль целостности и изменения внешних зависимостей. Для изолированного профиля дополнительно подтверждается перенос по разрешённой процедуре и отсутствие новой внешней рабочей зависимости.

Обновление сначала проходит в проверочном окружении с безопасными данными. Проверяются вход, права пространств, загрузка и поиск документов, защищённый модельный запрос, запрещённый маршрут, экспорт аудита и удаление. Набор выбирается по изменению: новая версия DLP требует контрольных примеров распознавания, изменение схемы данных — проверки восстановления и совместимости.

Перед рабочим переключением сохраняются применённая конфигурация и пригодная точка восстановления. Если миграция необратима, возврат старого образа недостаточен: заранее определяется способ восстановления данных и допустимая потеря записей после переключения. Поэтапный ввод и ограниченная проверка относятся к процедуре эксплуатации; конкретный механизм зависит от топологии поставки.

После обновления сравниваются ошибки, задержки, решения Guardrail и доставка событий с исходным состоянием. Ухудшение разбирается по стадии и затронутому сценарию. Отсутствие ошибок запуска не подтверждает корректную изоляцию данных или сохранение прежней политики.

Приёмочный протокол

Протокол содержит идентификатор окружения, версии, фактическую топологию, время испытания, выполненный отказ, использованную копию или точку, измеренные RPO/RTO и результаты проверки прав. Для базового профиля ежемесячное восстановление подтверждается таким протоколом, а не только уведомлением «копирование завершено».

Заказчик проверяет три результата: сервис восстановлен в установленной области, данные соответствуют выбранной точке и защитные ограничения сохранились. Отдельно фиксируются непроверенные отказы и несовпадения с целевым профилем. Они становятся конкретными эксплуатационными действиями с владельцами; до повторного испытания соответствующий показатель остаётся неподтверждённым.