
DATAAAD

Импорт и объединение датасетов

Полное руководство пользователя

01 **Импорт**

Файл становится датасетом

02 **Preview**

Точные счётчики и риски

03 **Результат**

Новый самостоятельный датасет

Union All

объединение всех строк

Union Distinct

объединение уникальных строк

Intersect

пересечение

Except
разность

Reconcile
сверка

Join Columns
присоединение столбцов

Сопоставление - составные ключи - предпросмотр - происхождение - Auto Parts

Содержание

Руководство построено от базового рабочего процесса к ключевым операциям, диагностике и API.

01 Главное за минуту	11 Intersect и Except (общие и отсутствующие строки)
02 Новые возможности сверки	12 Reconcile (расширенная сверка)
03 Быстрый выбор операции	13 Join Columns (присоединение столбцов по ключу)
04 Что означают операции	14 Saved Rules (сохранённые правила)
05 Другие термины в руководстве	15 Preview (предпросмотр), который нельзя обойти
06 Базовый сценарий	16 Результат, Lineage (происхождение) и Metadata (метаданные)
07 Какие строки участвуют: условия И/ИЛИ	17 Практические рецепты
08 Основная таблица и источники	18 Troubleshooting (что делать при проблемах)
09 Mapping (сопоставление столбцов) для вертикального объединения	19 Примеры на Auto Parts
10 Key (ключ): как DATAAD понимает совпадение	

Пользовательское руководство по вкладкам «Импорт» и «Объединение».
Обновлено 3 августа 2026 года.
PDF-версия: [datalad-dataset-combine-guide-ru.pdf](#).

Главное за минуту

DATA Lad не дописывает CSV или Excel прямо внутрь существующего датасета. Рабочий процесс состоит из двух безопасных шагов:

```
файл CSV/XLSX
→ Импорт
отдельный проверяемый датасет
→ Объединение с другими датасетами
новый датасет
```

Исходные датасеты не меняются. Результат всегда создаётся отдельно, получает связи с родителями и может быть открыт в гриде.

Перед каждым запуском сервер строит точный предпросмотр (preview). Запускается именно проверенная конфигурация: если режим, связи столбцов, ключи или входные данные изменились, DATA Lad требует новый предпросмотр.

Новые возможности сверки

В этой редакции усилен режим **Сверить**:

- над настройкой постоянно показано, **по каким полям ищется запись и какие значения сравниваются**;
- если обычное проверяемое поле по ошибке добавлено в составной ключ и покрытие упало, preview называет эту пару и предлагает одной кнопкой перенести её в сравниваемые значения;
- текстовые сравниваемые значения можно сверять без учёта регистра, при этом ключи остаются строгими;
- результат содержит [different_fields](#) и [difference_count](#), поэтому для каждой строки видно, какие именно столбцы разошлись;
- сохранённое правило восстанавливает режим, ключи, сравниваемые значения и настройку регистра в тех же ролях.

Ключ по-прежнему отвечает только на вопрос «это та же запись?». Опечатка в ключе не превращается в предполагаемое совпадение: строка честно получает статус «только в основной», а похожие ключи можно проверить в диагностических примерах.

Быстрый выбор операции

Устоявшийся английский термин	Русский перевод и смысл	Подпись в интерфейсе
Union All	Объединение всех строк: вертикально сложить таблицы и ничего не удалять	Добавить все строки

Устоявшийся английский термин	Русский перевод и смысл	Подпись в интерфейсе
Union Distinct	Объединение уникальных строк: вертикально сложить таблицы и убрать полные дубли	Убрать полные дубликаты
Intersect	Пересечение: оставить строки основной таблицы, ключи которых есть во всех источниках	Только общие строки
Except	Разность: вычесть из основной таблицы ключи источников	Только отсутствующие в источнике
Reconcile	Сверка: разделить строки на совпавшие, изменённые и отсутствующие	Сверить
Join Columns	Присоединение столбцов: обогатить основную таблицу полями источника по ключу	Присоединить столбцы по ключу

Если задача начинается с файла, сначала используйте «Импорт». Во вкладке «Объединение» выбираются только уже зарегистрированные датасеты.

Что означают операции

Union All (объединение всех строк)

Это вертикальное объединение: строки источника помещаются под строками основной таблицы. Ни одна строка не удаляется, даже если она полностью совпадает с уже существующей.

Пример:

```
Свод:    1000 строк
Поставка: 50 строк
Результат: 1050 строк
```

Используйте эту операцию для журналов, транзакций и ежедневных поставок, когда важно сохранить каждый полученный экземпляр строки.

Union Distinct (объединение уникальных строк)

Сначала DATAΛAD складывает строки вертикально, затем оставляет только один экземпляр каждой полностью одинаковой строки. Сравнивается вся результирующая строка, а не только идентификатор.

Две записи с одинаковым **sku**, но разными ценами не являются полными дублями и обе останутся. Операция подходит для объединения выгрузок, которые могут частично перекрывать друг друга.

Intersect (пересечение по ключу)

Операция отвечает на вопрос: «Какие строки основной таблицы присутствуют во всех выбранных источниках?»

DATAΛAD сравнивает не всю строку, а выбранный простой или составной ключ. В результат попадают полные строки основной таблицы. Строки и дополнительные столбцы источников не копируются.

Эксперт (разность по ключу)

Операция отвечает на вопрос: «Какие строки основной таблицы не встречаются ни в одном из выбранных источников?»

Для одного источника это обычная направленная разность:

```
Primary - Source  
A - B
```

Для нескольких источников из основной таблицы вычитается объединение их ключей:

```
A - (B ∪ C ∪ ...)
```

Это удобно для поиска новых записей, пропусков, расхождений между реестрами и объектов, которые ещё не были обработаны.

Reconcile (сверка по ключу и значениям)

Операция не отбрасывает строки основной таблицы, а классифицирует их:

- ключ найден и значения совпадают;
- ключ найден, но выбранные значения изменились;
- ключ отсутствует в источнике;
- ключ существует только в источнике - это показывается в диагностике.

Результатом становится отдельный отчёт со статусами и указанием, в каких источниках найден или не найден каждый ключ.

Join Columns (присоединение столбцов по ключу)

Операция обогащает строки основной таблицы столбцами источника. Для каждого ключа основной таблицы DATA/LAD ищет одну соответствующую строку источника и переносит её дополнительные значения.

Это аналог левого соединения таблиц: все строки основной таблицы сохраняются, а при отсутствии совпадения новые столбцы остаются пустыми. Источник должен быть уникален по ключу, иначе одна строка основной таблицы неоднозначно соответствовала бы нескольким строкам источника.

Другие термины в руководстве

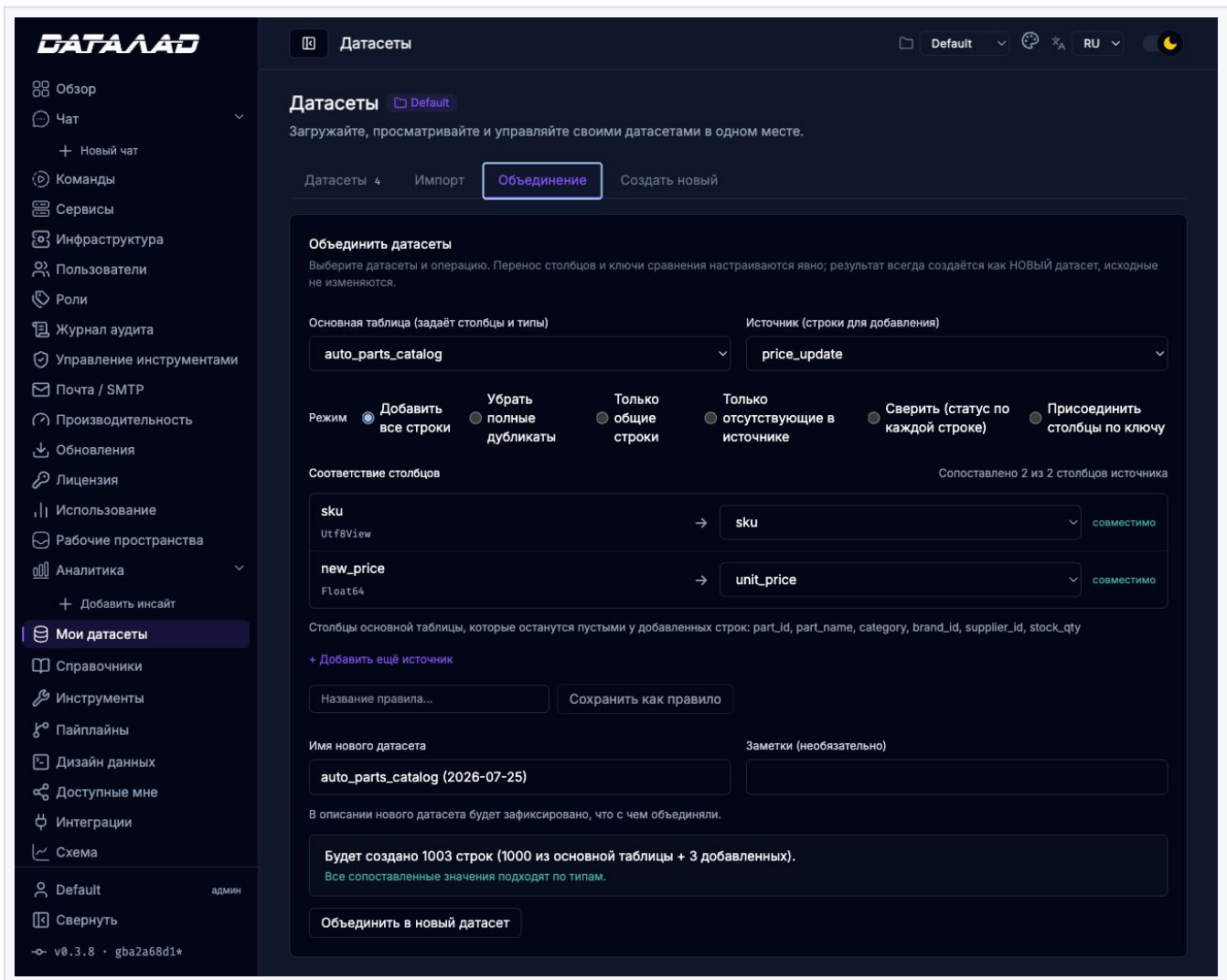
Профессиональные английские названия сохранены. При первом употреблении рядом даётся русский смысл:

Термин	Пояснение
Primary (основная таблица)	Таблица, которая задаёт строки или схему результата
Source (источник)	Дополнительный датасет, который добавляет строки, участвует в сравнении или отдаёт столбцы
Mapping (сопоставление столбцов)	Связь вида <code>source_column → primary_column</code>

Термин	Пояснение
Key (ключ)	Один столбец или набор столбцов, идентифицирующих запись
Composite Key (составной ключ)	Ключ из нескольких полей, сравниваемых вместе
Preview (предпросмотр)	Точный расчёт результата, потерь и проблем до запуска
Key Quality (качество ключа)	Уникальность, пустые значения, покрытие, дубли и кардинальность
Cardinality (кардинальность)	Отношение числа строк на один ключ: 1:1 , N:1 , 1:N , N:N
Commit (фиксация результата)	Создание нового датасета строго по проверенному Preview
Lineage (происхождение данных)	Связи результата с основной таблицей, источниками и их версиями
Metadata (метаданные)	Формулы, форматы отображения, словари и виртуальные связи

Базовый сценарий

- 1 Откройте **Мои датасеты** → **Импорт**.
- 2 Импортируйте CSV или XLSX как отдельный датасет.
- 3 Откройте импортированный датасет и убедитесь, что заголовки, типы и строки распознаны правильно.
- 4 Перейдите во вкладку **Объединение**.
- 5 Выберите основную таблицу. Она задаёт схему, типы, рабочее пространство и семейство результата.
- 6 Выберите источник.
- 7 Выберите режим.
- 8 Проверьте сопоставление столбцов или ключи.
- 9 При необходимости раскройте **Какие строки участвуют** и задайте отдельные группы И/ИЛИ для основной таблицы и каждого источника.
- 10 Дождитесь предпросмотра и изучите воронку «было → прошло условия», счётчики результата, потери преобразования и качество ключей.
- 11 Задайте имя и заметки результата.
- 12 Нажмите **Объединить в новый датасет**.
- 13 Откройте результат в гриде.



Сопоставление столбцов и предпросмотр добавления строк

На примере выше `price_update.sku` переносится в `auto_parts_catalog.sku`, а `new_price` - в `unit_price`. Предпросмотр заранее сообщает: будет 1003 строки, типы совместимы, остальные столбцы каталога у трёх добавленных строк будут пустыми.

Какие строки участвуют: условия И/ИЛИ

Раскройте этот блок, если в операции нужен не весь датасет. Основная таблица и каждый источник настраиваются независимо. Например:

- в основной таблице: **Статус = Активен**;
- в источнике: **(Регион = UAE И Дата >= 2026-01-01) ИЛИ Приоритет = Высокий**.

Сначала DATAAD применяет эти условия к каждому входу, затем выполняет mapping, поиск по ключу или сверку значений. Поэтому условие не становится частью составного ключа и не меняет смысл **different**.

В Preview у каждого входа показывается воронка, например **120 000 → 18 430** и число исключённых строк. Следом показывается уже результат выбранного режима. Условия входят в обязательный Preview и в сохранённое правило: изменение любого поля, оператора, значения или И/ИЛИ требует нового расчёта.

DATA LAB

Датасеты

Основное рабочее пространство

Датасеты 2 Импорт **Объединение** Создать новый

Объединить датасеты

Выберите датасеты и операцию. Перенос столбцов и ключи сравнения настраиваются явно; результат всегда создается как **НОВЫЙ** датасет, исходные не изменяются.

Основная таблица (задаёт столбцы и типы) Источник (строки для добавления)

Каталог автозапчастей Обновление цен и остатков

Режим **Добавить все строки** Убрать полные дубликаты Только общие строки Только отсутствующие в источнике Сверить (статус по каждой строке) Присоединить столбцы по ключу

Какие строки участвуют
Основная таблица: условий 1; источник: 3

Основная таблица
Оставьте только строки основной таблицы, которые должны войти в операцию.

Группа 1 Все Любое

Статус равно Активен

+ Добавить условие

+ Добавить группу

Источник
Оставьте только строки этого источника до сопоставления и объединения.

Все группы Любая группа

Группа 1 Все Любое

Регион равно UAE

Дата обновления не меньше 01/01/2026

+ Добавить условие

Группа 2 Все Любое

Приоритет равно Высокий

+ Добавить условие

+ Добавить группу

Сопоставление столбцов

Сопоставлено 2 из 5 столбцов источника

Столбец основной таблицы	Столбец источника	Статус
Регион	— не переносить	игнорируется
Дата обновления	— не переносить	игнорируется
Приоритет	— не переносить	игнорируется
Артикул	Артикул	совместимо
Новая цена	Цена	совместимо

Столбцы основной таблицы, которые останутся пустыми у добавленных строк: Статус, Регион

+ Добавить ещё источник

Название правила... Сохранить как правило

Имя нового датасета Заметки (необязательно)

Каталог автозапчастей (2026-08-03)

В описании нового датасета будет зафиксировано, что с чем объединяли.

Каталог автозапчастей 120000 → 18430 **исключено: 101570** Обновление цен и остатков 8200 → 2460 **исключено: 5740**

Будет создано 20890 строк (18430 из основной таблицы + 2460 добавленных).
Все сопоставленные значения подходят по типам.

Объединить в новый датасет

Администратор админ

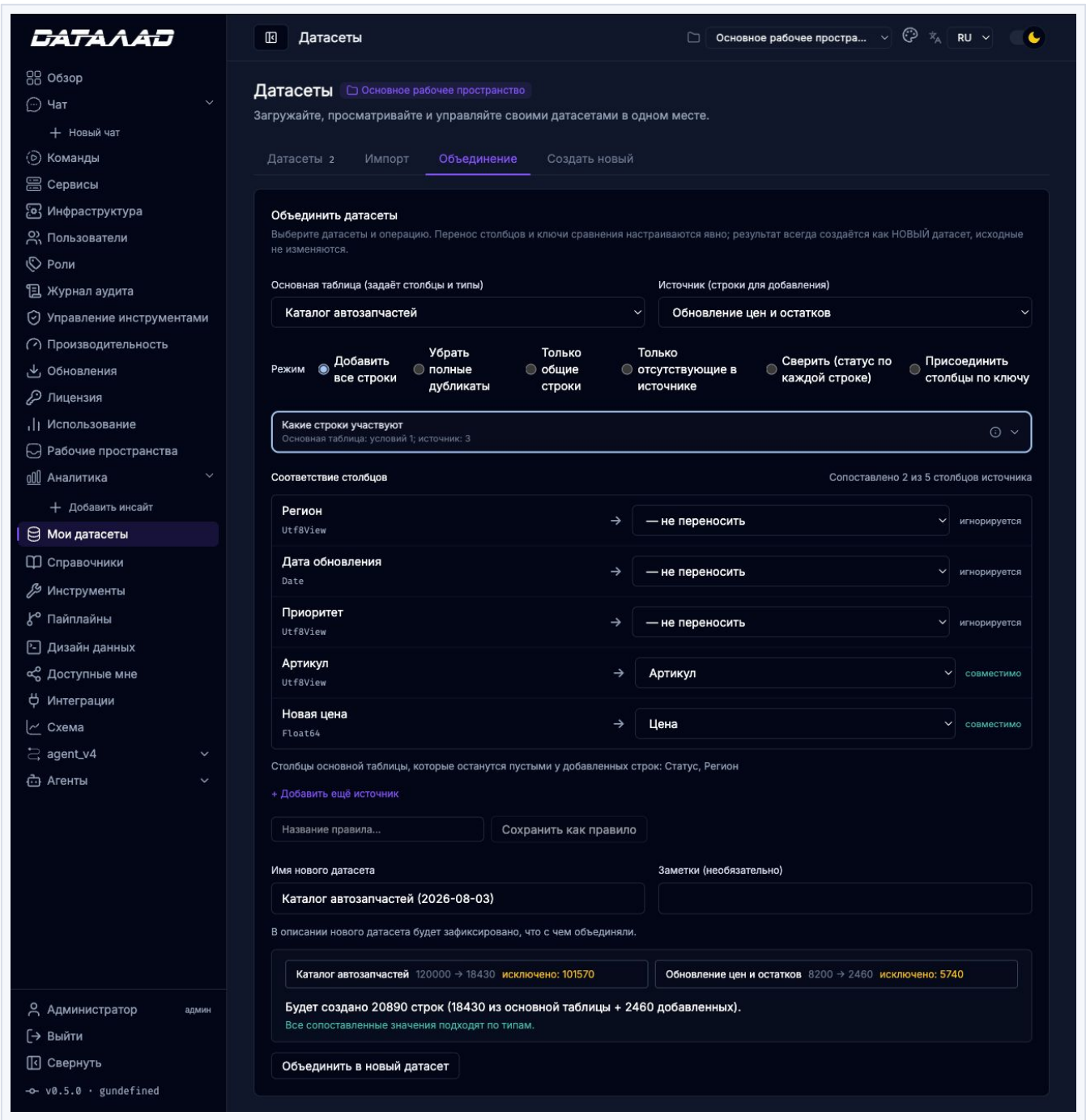
Выйти

Свернуть

v0.5.0 · gundefined

Независимые условия И/ИЛИ для основной таблицы и источника

На снимке основная таблица пропускает только активные записи. Источник пропускает строки по выражению **(Регион = UAE И Дата обновления >= 2026-01-01) ИЛИ Приоритет = Высокий**. Переключатели **Все / Любое** внутри карточки управляют условиями одной группы, а **Все группы / Любая группа** - связью между группами.



Preview с воронкой строк после условий

В свёрнутом состоянии блок занимает одну строку и показывает число условий. Ниже Preview отдельно фиксирует воронку каждого входа: сколько строк было, сколько прошло и сколько исключено. Итоговый счётчик рассчитывается уже по отфильтрованным наборам.

Основная таблица и источники

Основная таблица

Основная таблица определяет:

- имена и порядок столбцов при вертикальном объединении;
- целевые типы данных;

- набор исходных строк для фильтрации, сверки и join;
- рабочее пространство нового датасета;
- семейство и основного родителя результата;
- формулы, форматы, словари и виртуальные связи, которые нужно переиздать.

Выбор основной таблицы важен. Например:

- `parts_catalog` как основная и `brands` как источник даёт каталог деталей с атрибутами бренда;
- `orders` как основная и `parts_catalog` как источник даёт строки заказов с атрибутами детали;
- обратный `parts_catalog + orders` пытается присоединить много заказов к одной детали и поэтому блокируется безопасным join.

Источники

Для большинства режимов можно добавить несколько источников. Их порядок сохраняется в операции и provenance.

[Присоединить столбцы по ключу](#) сейчас принимает ровно один источник. Это сознательное ограничение безопасной первой версии materialized join.

Mapping (сопоставление столбцов) для вертикального объединения

Режимы **Добавить все строки** и **Убрать полные дубликаты** используют связи:

столбец источника → столбец основной таблицы

DATA2DAD предлагает связи детерминированно: по точному и нормализованному имени, похожести имени, совместимости типов и последнему успешному запуску. Подсказка всегда редактируема и сама по себе ничего не запускает.

Правила Mapping:

- один столбец основной таблицы нельзя занять дважды в рамках одного источника;
- несопоставленные столбцы источника игнорируются;
- несопоставленные столбцы основной таблицы получают `NULL` в добавляемых строках;
- значения приводятся к типу основной таблицы;
- невозможные преобразования становятся `NULL`, а их количество показывается как потеря преобразования до запуска.

Добавить все строки

Результат идёт в порядке:

все строки основной таблицы
→ все строки источника 1
→ все строки источника 2
→ ...

Полные дубли сохраняются. Это основной режим ежедневной дозагрузки, когда история поставок должна остаться целиком.

Убрать полные дубликаты

После выравнивания столбцов удаляются полностью одинаковые output-строки. Preview показывает, сколько экземпляров будет удалено.

Это не дедупликация по [sku](#), [order_id](#) или другому бизнес-ключу. Если две строки имеют один ключ, но отличаются хотя бы одним значением, обе останутся. Порядок строк после [DISTINCT](#) не гарантируется.

Key (ключ): как DATALAD понимает совпадение

Режимы **Только общие**, **Только отсутствующие**, **Сверить** и **Присоединить столбцы** используют отдельный конструктор ключа.

Простой ключ:

```
source.sku → primary.sku
```

Составной ключ:

```
source.order_no → primary.order_id  
source.company → primary.company_id
```

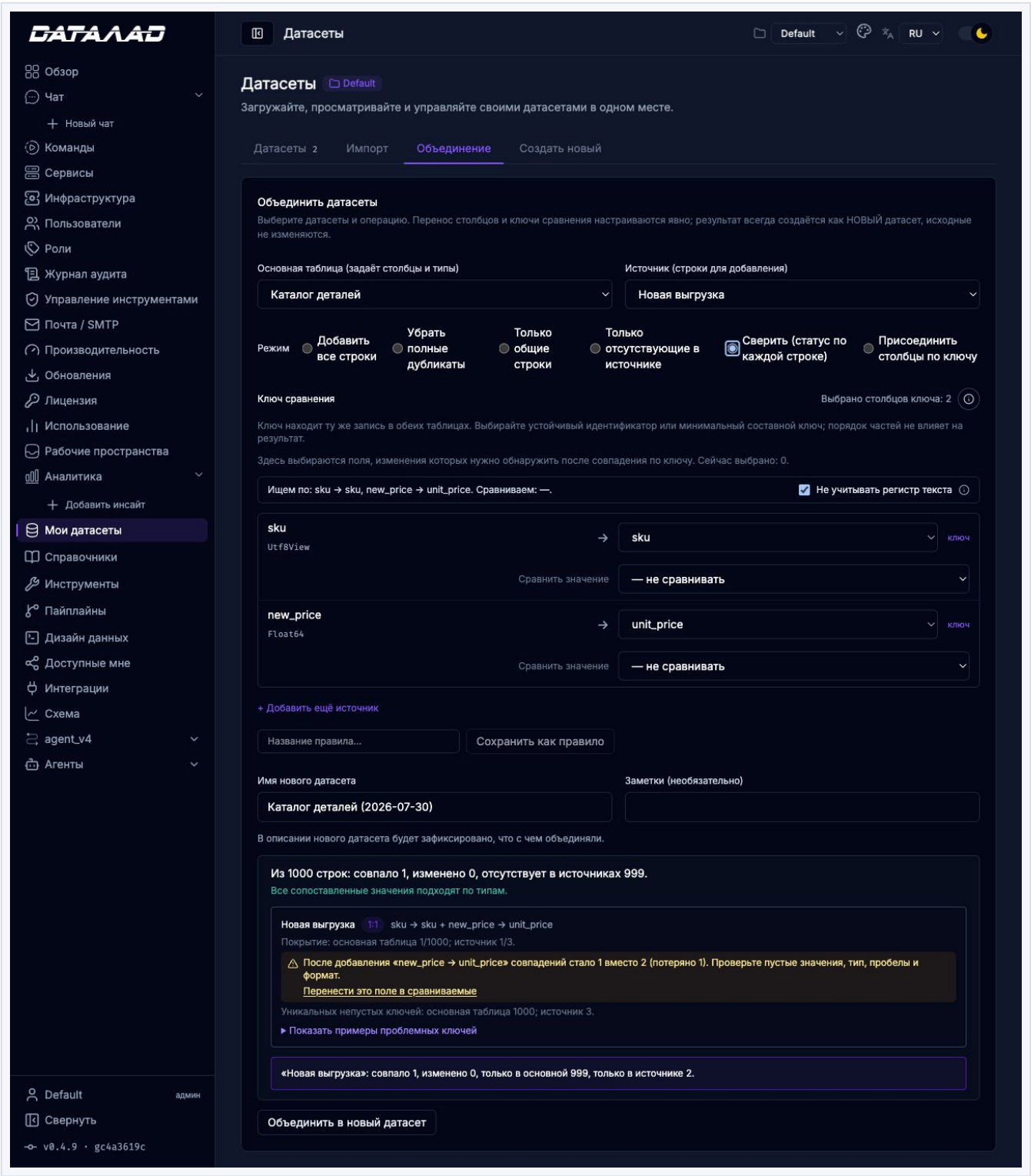
Части ключа сравниваются вместе. Порядок, в котором они были добавлены, не влияет на итог: [sku + company](#) и [company + sku](#) задают то же условие при тех же парах столбцов. Значения источника сначала приводятся к типам соответствующих столбцов основной таблицы. Пустой хотя бы в одной части ключ считается пустым целиком и не совпадает с другим [NULL](#).

Ключ должен быть минимальным устойчивым идентификатором записи. Не включайте в него цену, статус, описание и другие поля только ради того, чтобы проверить их равенство. Для этого в режиме **Сверить** предназначен отдельный выбор **Сравнить значение**.

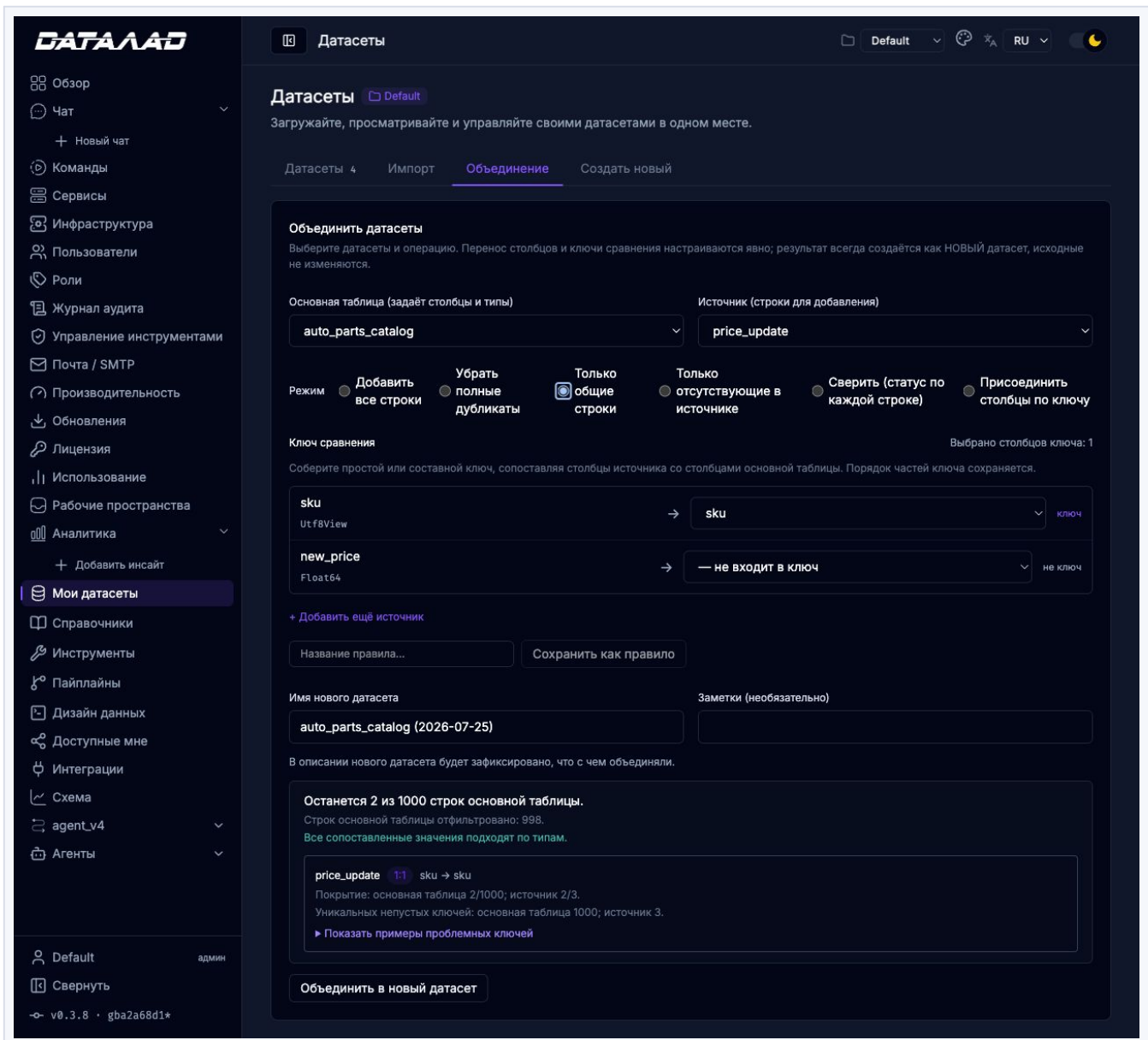
Если новая часть составного ключа уменьшила покрытие, интерфейс показывает точную добавленную пару и изменение счётчика: например, «совпадений стало 0 вместо 367 817». В такой ситуации проверьте пустые значения, типы, невидимые пробелы, переносы строк и формат даты или числа.

Если покрытие упало после новой пары

- 1 Посмотрите имя пары в жёлтом предупреждении.
- 2 Раскройте **Показать примеры проблемных ключей**.
- 3 Проверьте счётчик пустых ключей и потери преобразования типов.
- 4 Если поле описывает состояние записи, а не её идентичность, уберите его из ключа и выберите в **Сравнить значение**.
- 5 Повторите preview и убедитесь, что покрытие вернулось, а реальные расхождения появились в категории **изменено**.



Предупреждение о лишнем поле в ключе и перенос в сравниваемые значения



Preview качества ключа

В preview отображаются:

- покрытие строк основной таблицы и источника;
- число уникальных непустых ключей;
- число строк с пустыми ключами;
- лишние строки-дубли;
- максимальный размер группы одного ключа;
- кардинальность;
- до пяти примеров для каждой проблемной категории.

Cardinality (кардинальность)

Значок	Значение	Практический смысл
1:1	ключ уникален с обеих сторон	идеальная сверка или join
N:1	ключ повторяется в основной таблице, уникален в источнике	безопасное обогащение справочником
1:N	ключ уникален в основной таблице, повторяется в источнике	join размножил бы строки основной таблицы
N:N	ключ повторяется с обеих сторон	неоднозначное сопоставление

Для materialized join источник обязан быть уникален по ключу. 1:N и N:N блокируют запуск. Повторы в основной таблице допустимы: один справочный атрибут может применяться к нескольким строкам.

Problem Key Examples (примеры проблемных ключей)

Раскройте **Показать примеры проблемных ключей**, чтобы увидеть конкретные значения:

- пустые ключи основной таблицы;
- пустые ключи источника;
- повторяющиеся ключи;
- ключи только в основной таблице;
- ключи только в источнике.

Объединить датасеты

Выберите датасеты и операцию. Перенос столбцов и ключи сравнения настраиваются явно; результат всегда создаётся как НОВЫЙ датасет, исходные не изменяются.

Основная таблица (задаёт столбцы и типы) **auto_parts_catalog** Источник (строки для добавления) **price_update**

Режим: ☒ Добавить все строки ☐ Убрать полные дубликаты ☒ Только общие строки ☐ Только отсутствующие в источнике ☐ Сверить (статус по каждой строке) ☐ Присоединить столбцы по ключу

Ключ сравнения Выбрано столбцов ключа: 1

Соберите простой или составной ключ, сопоставляя столбцы источника со столбцами основной таблицы. Порядок частей ключа сохраняется.

sku Utf8View	→	sku	ключ
new_price Float64	→	— не входит в ключ	не ключ

+ Добавить ещё источник

Название правила... Сохранить как правило

Имя нового датасета **auto_parts_catalog (2026-07-25)** Заметки (необязательно)

В описании нового датасета будет зафиксировано, что с чем объединяли.

Останется 2 из 1000 строк основной таблицы.
 Строк основной таблицы отфильтровано: 998.
 Все сопоставленные значения подходят по типам.

Показать примеры проблемных ключей

Только в основной таблице	Только в источнике
BDY-00002 × 1	UNKNOWN-999 × 1
BDY-00003 × 1	
BDY-00004 × 1	
BDY-00005 × 1	
BDY-00006 × 1	

Объединить в новый датасет

Конкретные unmatched key examples

Примеры ограничены пятью значениями на категорию и предназначены для быстрой диагностики. Счётчики при этом относятся ко всему датасету.

Intersect и Except (общие и отсутствующие строки)

Только общие строки

Результат содержит полные строки основной таблицы, ключ которых присутствует во всех выбранных источниках.

- строки источников в результат не добавляются;
- дубли основной таблицы сохраняются;

- для нескольких источников используется логика «есть во всех»;
- **NULL**-ключи не совпадают.

Пример: из 1000 деталей в **parts_catalog** только два **sku** присутствуют в **price_update**; результат содержит две исходные строки каталога.

Только отсутствующие в источнике

Это **Эксерпт (разность)**. Результат содержит полные строки основной таблицы, ключ которых отсутствует во всех выбранных источниках. Источники здесь действуют как единый список ключей, который вычитается из основной таблицы.

Пример: получить детали, для которых не пришло обновление цены, или клиентов, которых нет в контрольном реестре.

Как проверить, что B полностью содержит A

Пусть **A** - контрольный список, а **B** - полный реестр:

```
A: P001, P002
B: P001, P002, P003
```

Выберите **A** как **Primary (основную таблицу)**, **B** как **Source (источник)** и выполните **Эксерпт (разность) A - B**:

```
A - B = ∅
```

Пустой результат означает: каждый непустой выбранный ключ из **A** найден в **B**. Значит, по этому ключу **A** действительно является подмножеством **B**.

Теперь поменяйте направление:

```
B - A = P003
```

Результат покажет полные строки **B**, которые не входят в **A**. Направление важно: операция всегда вычитает Source из Primary.

Эта проверка доказывает включение именно **по выбранному ключу**. Если у **P001** в таблицах разные цены или названия, ключ всё равно считается найденным. Чтобы проверить и значения, используйте **Reconcile (сверку)**: задайте идентификатор как Key (ключ), а цену, статус и другие поля - как сравниваемые значения. Не превращайте все проверяемые поля в составной ключ: пустое значение в любой его части сделает запись найденной вместо статуса «изменено».

Reconcile (расширенная сверка)

Режим **Сверить** отделяет:

- 1 ключ - идентичность записи;
- 2 сравниваемые поля - значения, которые должны совпасть.

Кнопка **i** открывает справку о ролях и пустых значениях. Два пустых сравниваемых значения равны; пустая часть ключа не совпадает даже с пустой.

DATA LAB

Датасеты

Датасеты 2 Импорт **Объединение** Создать новый

Объединить датасеты
Выберите датасеты и операцию. Перенос столбцов и ключи сравнения настраиваются явно; результат всегда создаётся как НОВЫЙ датасет, исходные не изменяются.

Основная таблица (задаёт столбцы и типы) **Каталог деталей** Источник (строки для добавления) **Новая выгрузка**

Режим ☒ Добавить все строки ☐ Убрать полные дубликаты ☐ Только общие строки ☐ Только отсутствующие в источнике ☒ Сверить (статус по каждой строке) ☐ Присоединить столбцы по ключу

Ключ сравнения Выбрано столбцов ключа: 1

Ключ находит ту же запись в обеих таблицах. Выберите устойчивый идентификатор или минимальный составной ключ; порядок частей не влияет на результат.

Здесь выбираются поля, изменения которых нужно обнаружить после совпадения по ключу. Сейчас выбрано: 1.

Ищем по: sku → sku. Сравниваем: new_price → unit_price. ☒ Не учитывать регистр текста

Ключ и сверяемые значения — это разные роли

- Ключ отвечает на вопрос «это та же запись?». Обычно это ID, номер договора, артикул или минимальная комбинация таких полей.
- «Сравнить значение» отвечает на вопрос «что изменилось в найденной записи?». Сюда относятся цена, статус, дата, описание и другие проверяемые поля.
- Пустая часть ключа не совпадает даже с другой пустой частью. В сверяемых значениях два пустых значения считаются равными.
- Сохранённое правило запоминает режим, ключи и сверяемые поля отдельно и восстанавливает их в тех же ролях.

Пример: «ID → ID» — ключ; «Цена → Цена» и «Статус → Статус» — сверяемые значения.

sku → sku **ключ**
Сравнить значение **— не сравнивать**

new_price → **— не входит в ключ** **не ключ**
Сравнить значение **unit_price**

+ Добавить ещё источник

Название правила... Сохранить как правило

Имя нового датасета **Каталог деталей (2026-07-30)** Заметки (необязательно)

В описании нового датасета будет зафиксировано, что с чем объединяли.

Из 1000 строк: совпало 1, изменено 1, отсутствует в источниках 998.
Все сопоставленные значения подходят по типам.

Новая выгрузка 1:1 sku → sku
Покрытие: основная таблица 2/1000; источник 2/3.
Уникальных непустых ключей: основная таблица 1000; источник 3.
► Показать примеры проблемных ключей

«Новая выгрузка»: совпало 1, изменено 1, только в основной 998, только в источнике 1.
Сравниваемые поля: new_price → unit_price.

Примеры изменённых значений
Ключ: BDY-00001
unit_price: 215.76 → 999.0

В результате для каждой строки будут добавлены «different_fields» со списком расходящихся столбцов и «difference_count» с их количеством.

Объединить в новый датасет

Встроенная справка о ролях ключа и сравниваемых значений

Например:

ключ: price_update.sku → parts_catalog.sku
сравнение: price_update.new_price → parts_catalog.unit_price

Preview показывает:

- **совпало** - ключ найден и выбранные значения одинаковы;
- **изменено** - ключ найден, но хотя бы одно значение отличается;
- **только в основной** - ключа нет хотя бы в одном источнике;
- **только в источнике** - source key отсутствует в основной таблице;
- до пяти примеров изменённых значений.

Быстрая проверка настройки сверки

- хотя бы одна пара выбрана как ключ;
- ключ действительно идентифицирует запись, а не описывает её состояние;
- цена, статус, дата и описание выбраны как сравниваемые значения;
- поля с возможными пустыми значениями не добавлены в ключ без необходимости;
- preview показывает ожидаемое покрытие и понятные примеры расхождений;
- сохранённое правило после повторного выбора восстанавливает тот же режим, ключи и сравниваемые значения.

DATA LAB

Датасеты

Датасеты 4 Импорт **Объединение** Создать новый

Объединить датасеты
Выберите датасеты и операцию. Перенос столбцов и ключи сравнения настраиваются явно; результат всегда создаётся как НОВЫЙ датасет, исходные не изменяются.

Основная таблица (задаёт столбцы и типы) **auto_parts_catalog** Источник (строки для добавления) **price_update**

Режим: ☒ Добавить все строки ☐ Убрать полные дубликаты ☐ Только общие строки ☐ Только отсутствующие в источнике ☒ Сверить (статус по каждой строке) ☐ Присоединить столбцы по ключу

Ключ сравнения: **sku** (Utf8View) → **sku** (Utf8View) Сравнить значение: **— не сравнивать** Выбрано столбцов ключа: 1

Соберите простой или составной ключ, сопоставляя столбцы источника со столбцами основной таблицы. Порядок частей ключа сохраняется. Для сверки можно дополнительно выбрать сравниваемые значения. Сейчас выбрано: 1.

new_price (Float64) → **— не входит в ключ** Сравнить значение: **unit_price** не ключ

+ Добавить ещё источник

Название правила... Сохранить как правило

Имя нового датасета: **auto_parts_catalog (2026-07-25)** Заметки (необязательно)

В описании нового датасета будет зафиксировано, что с чем объединяли.

Из 1000 строк: совпало 1, изменено 1, отсутствует в источниках 998.
Все сопоставленные значения подходят по типам.

price_update 1:1 sku → sku
Покрытие: основная таблица 2/1000; источник 2/3.
Уникальных непустых ключей: основная таблица 1000; источник 3.
► Показать примеры проблемных ключей

«price_update»: совпало 1, изменено 1, только в основной 998, только в источнике 1.
Сравниваемые поля: new_price → unit_price.
Примеры изменённых значений
Ключ: BDY-00001
unit_price: 215.76 → 999.0

Объединить в новый датасет

Сверка цены с примером изменённого значения

В Auto Parts примере:

- **SUS-00001**: цена совпала;
- **BDY-00001**: 215.76 → 999.0;
- 998 деталей отсутствуют в маленькой поставке;
- **UNKNOWN-999** существует только в поставке.

Новый reconcile-датасет сохраняет grain основной таблицы и добавляет:

Столбец	Значение
match_status	matched, different или only_primary
present_in	id источников, где найден ключ
missing_in	id источников, где ключ не найден
different_fields	столбцы с расхождениями; пусто, если сравнить не с чем
difference_count	количество точно найденных расхождений

Строки, существующие только в источнике, не добавляются в результат. Их количество и примеры остаются в preview. Если при сравнении значений источник имеет несколько строк на один ключ, запуск блокируется: невозможно однозначно выбрать сравниваемое значение.

Для текстовых сверяемых значений можно включить **Не учитывать регистр текста**. Например, Dubai и DUBAI тогда равны. Настройка не применяется к ключам: опечатка или другой регистр идентификатора по-прежнему означает, что точного кандидата нет.

DATA LAB

Датасеты

Датасеты 2 | Импорт | **Объединение** | Создать новый

Объединить датасеты
Выберите датасеты и операцию. Перенос столбцов и ключи сравнения настраиваются явно; результат всегда создаётся как НОВЫЙ датасет, исходные не изменяются.

Основная таблица (задаёт столбцы и типы) | Источник (строки для добавления)
Каталог деталей | Новая выгрузка

Режим: **Добавить все строки** | Убрать полные дубликаты | Только общие строки | Только отсутствующие в источнике | Сверить (статус по каждой строке) | Присоединить столбцы по ключу

Ключ сравнения | Выбрано столбцов ключа: 1

Ключ находится ту же запись в обеих таблицах. Выберите устойчивый идентификатор или минимальный составной ключ; порядок частей не влияет на результат.
Здесь выбираются поля, изменения которых нужно обнаружить после совпадения по ключу. Сейчас выбрано: 1.

Ищем по: sku → sku. Сравниваем: new_price → unit_price. ☒ Не учитывать регистр текста

Поле	Тип	Ключ	Сравнить значение
sku	Utf8View	sku	— не сравнивать
new_price	Float64	— не входит в ключ	unit_price

+ Добавить ещё источник

Название правила... | Сохранить как правило

Имя нового датасета | Заметки (необязательно)
Каталог деталей (2026-07-30)

В описании нового датасета будет зафиксировано, что с чем объединяли.

Из 1000 строк: совпало 1, изменено 1, отсутствует в источниках 998.
Все сопоставленные значения подходят по типам.

Новая выгрузка 1:1 sku → sku
Покрытие: основная таблица 2/1000; источник 2/3.
Уникальных непустых ключей: основная таблица 1000; источник 3.
► Показать примеры проблемных ключей

«Новая выгрузка»: совпало 1, изменено 1, только в основной 998, только в источнике 1.
Сравниваемые поля: new_price → unit_price.
Примеры изменённых значений
Ключ: BDY-00001
unit_price: 215.76 → 999.0
В результате для каждой строки будут добавлены «different_fields» со списком расходящихся столбцов и «difference_count» с их количеством.

Объединить в новый датасет

Default админ
Свернуть
v0.4.9 · gc4a3619c

Сводка ролей, сравнение без регистра и поля результата сверки

Join Columns (присоединение столбцов по ключу)

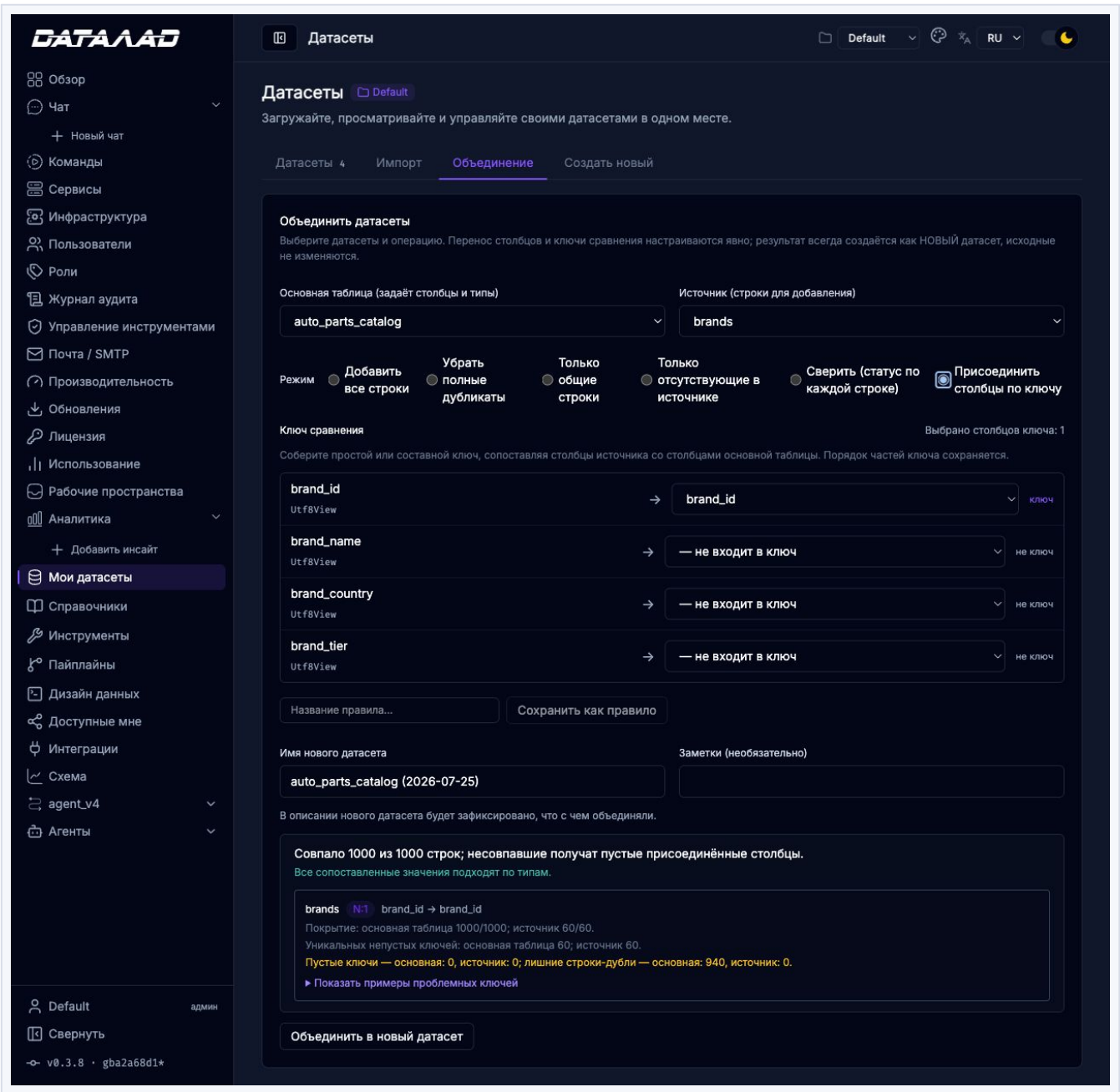
Это materialized **LEFT JOIN**:

- все строки основной таблицы сохраняются;

- все non-key столбцы источника добавляются справа;
- отсутствующее совпадение даёт **NULL** в новых столбцах;
- коллизии имён получают суффиксы **_2**, **_3**, ...;
- источник должен быть уникален по ключу.

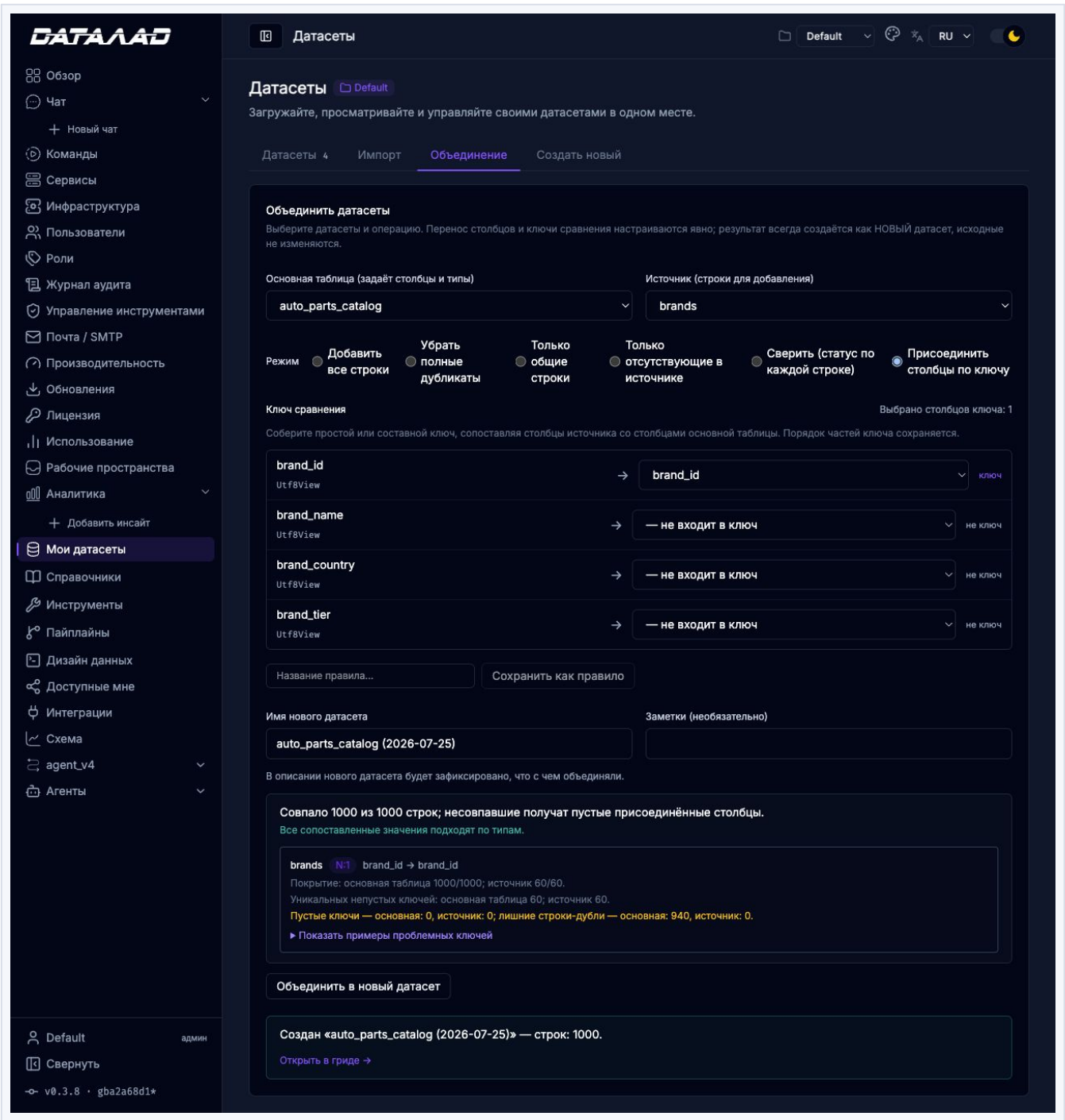
Безопасный сценарий: каталог + справочник

`parts_catalog.brand_id` → `brands.brand_id` имеет кардинальность **N:1**: в каталоге один бренд встречается у многих деталей, но в справочнике каждый бренд представлен одной строкой.



Безопасный N:1 join со справочником брендов

Preview показывает покрытие 1000/1000 и 60 уникальных ключей справочника. После запуска создаётся новый датасет на 1000 строк.



Успешно созданный новый датасет

Небезопасный сценарий: каталог + заказы

В `orders` один `part_id` встречается много раз. Попытка присоединить заказы к каталогу размножила бы одну деталь на несколько строк. DATA LAB показывает `1:N`, число повторов и блокирует кнопку.

Датасеты Default

Загружайте, просматривайте и управляйте своими датасетами в одном месте.

Датасеты 4 Импорт **Объединение** Создать новый

Объединить датасеты
Выберите датасеты и операцию. Перенос столбцов и ключи сравнения настраиваются явно; результат всегда создаётся как НОВЫЙ датасет, исходные не изменяются.

Основная таблица (задаёт столбцы и типы) auto_parts_catalog Источник (строки для добавления) orders

Режим: ☒ Добавить все строки ☐ Убрать полные дубликаты ☐ Только общие строки ☐ Только отсутствующие в источнике ☐ Сверить (статус по каждой строке) ☒ Присоединить столбцы по ключу

Ключ сравнения Выбрано столбцов ключа: 1

Соберите простой или составной ключ, сопоставляя столбцы источника со столбцами основной таблицы. Порядок частей ключа сохраняется.

Столбец	Тип	Ключ	Статус
order_id	Utf8View	— не входит в ключ	не ключ
part_id	Utf8View	part_id	ключ
customer_id	Utf8View	— не входит в ключ	не ключ
order_date	Date32	— не входит в ключ	не ключ
quantity	Int64	— не входит в ключ	не ключ
unit_price	Float64	— не входит в ключ	не ключ
discount_pct	Int64	— не входит в ключ	не ключ
line_total	Float64	— не входит в ключ	не ключ
region	Utf8View	— не входит в ключ	не ключ
channel	Utf8View	— не входит в ключ	не ключ

Название правила... Сохранить как правило

Имя нового датасета auto_parts_catalog (2026-07-25) Заметки (необязательно)

В описании нового датасета будет зафиксировано, что с чем объединяли.

Совпало 779 из 1000 строк; несовпавшие получают пустые присоединённые столбцы.
Все сопоставленные значения подходят по типам.

```
orders 1:N part_id → part_id
```

Покрывие: основная таблица 779/1000; источник 1500/1500.
Уникальных непустых ключей: основная таблица 1000; источник 779.
Пустые ключи — основная: 0, источник: 0; лишние строки-дубли — основная: 0, источник: 721.
[► Показать примеры проблемных ключей](#)

Соединение невозможно: в источнике «orders» строк-дублей ключа: 721. Добавьте части ключа или удалите дубли в источнике.

Join заблокирован из-за повторяющихся ключей источника

В демо-данных источник содержит 1500 заказов, 779 уникальных `part_id` и 721 лишнюю строку-дубли ключа.

Правильные варианты зависят от задачи:

- если нужна таблица заказов с названием детали, сделайте `orders` основной таблицей, а уникальный `parts_catalog` - источником;

- если нужна одна строка на деталь, сначала агрегируйте заказы до grain `part_id`, затем присоедините агрегат;
- если реальный ключ составной, добавьте недостающие части ключа;
- если повторы ошибочны, очистите источник и повторите preview.

DATAŁAD намеренно не выбирает «первую» или «последнюю» строку скрыто.

Saved Rules (сохранённые правила)

Когда одинаковая поставка приходит регулярно:

- 1 настройте режим, mapping либо ключи и сравниваемые значения;
- 2 введите название правила;
- 3 нажмите **Сохранить как правило**;
- 4 в следующий раз выберите новый датасет поставки и примените правило.

Правило привязано к семейству основной таблицы, а не к конкретной ежедневной версии. Поэтому оно продолжает работать после создания следующего «Свода».

Правило запоминает роли отдельно:

- выбранный режим операции;
- mapping для вертикального объединения;
- ключи идентичности записи;
- сравниваемые значения режима **Сверить**;
- политику сравнения текста с учётом или без учёта регистра.
- условия участия строк основной таблицы и источника, включая группы И/ИЛИ.

При повторном применении цена или статус, сохранённые как сравниваемые значения, не превращаются в части ключа. Старые правила, созданные до появления отдельных ролей и политики регистра, продолжают открываться как раньше, используют точное сравнение текста и не требуют ручной миграции.

При применении DATAŁAD проверяет свежую схему и сообщает:

- сколько связей применено;
- какой столбец исчез в источнике;
- какой целевой столбец исчез в основной таблице.

Частичное применение вида «N из M» не скрывается. Перед запуском оператор видит, какие связи требуют исправления. Последний успешный запуск также используется для подсказки «как в прошлый раз».

Preview (предпросмотр), который нельзя обойти

Preview - обязательный точный расчёт, а не приблизительная подсказка.

Он фиксирует:

- пользователя;
- режим;
- основную таблицу и её конкретную версию;
- упорядоченные источники и их версии;
- fingerprints схем;
- условия участия primary и каждого source;
- mapping;
- ключи;
- сравниваемые поля;
- результат blocking checks.

После Preview запуск принимает только ту конфигурацию данных, которую пользователь только что проверил.

Новый preview потребуется, если:

- изменён primary или source;
- добавлен, удалён или переставлен источник;
- изменён режим;
- изменён фильтр, mapping, ключ, compare или политика сравнения регистра;
- один из входных датасетов получил новую текущую версию;
- приложение было перезапущено.

Имя и заметки результата можно менять без нового preview: они не меняют данные. При конфликте ничего не регистрируется, а временный результат удаляется.

Результат, Lineage (происхождение) и Metadata (метаданные)

Каждый результат:

- является новым самостоятельным датасетом;
- продолжает семейство основной таблицы;
- хранит каждого родителя, его роль и pinned version;
- получает provenance с режимом и настройками;
- не изменяет входы.

Из основной таблицы переиздаются:

- формулы столбцов;
- display formats;
- dictionary bindings;
- Grid Merge definitions.

Виртуальные столбцы не запекаются в Parquet. Ошибка переиздания отдельного вида metadata не уничтожает уже созданные данные: она возвращается как предупреждение. Комментарии, стили и tags отдельных строк/ячеек не переносятся; результат начинает новую row history.

Практические рецепты

Ежедневно добавить поставку в «Свод»

- 1 Импортировать файл как [Поставка 25.07](#).
- 2 Primary: текущий [Свод](#).
- 3 Source: [Поставка 25.07](#).
- 4 Режим: **Добавить все строки**.
- 5 Применить сохранённое правило.
- 6 Проверить несопоставленные столбцы и [cast_nulls](#).
- 7 Создать [Свод 25.07](#).

Найти новые записи

- 1 Primary: новая поставка.
- 2 Source: текущий реестр.
- 3 Режим: **Только отсутствующие в источнике**.
- 4 Ключ: бизнес-идентификатор записи.

Результат содержит строки новой поставки, которых ещё нет в реестре.

Доказать, что A является подмножеством B

- 1 Primary: контрольный список [A](#).
- 2 Source: полный реестр [B](#).
- 3 Режим: **Экспорт (разность) - Только отсутствующие в источнике**.
- 4 Выбрать бизнес-ключ.
- 5 Проверить Preview и создать результат [A - B](#).
- 6 Убедиться, что результат пуст.

Пустой [A - B](#) означает, что все непустые ключи [A](#) присутствуют в [B](#). Чтобы увидеть записи, которые есть только в большем реестре, поменяйте датасеты местами и выполните [B - A](#).

Проверить, что обязательные записи присутствуют

- 1 Primary: контрольный список.
- 2 Добавить проверяемые источники.
- 3 Режим: **Только общие строки** для полного пересечения или **Только отсутствующие** для списка проблем.
- 4 Проверить coverage и unmatched examples.

Сверить изменения справочника

- 1 Primary: утверждённый справочник.

- 2 Source: новая версия.
- 3 Режим: **Сверить**.
- 4 Выбрать ключ.
- 5 Выбрать поля, которые должны совпасть.
- 6 Проверить changed samples и source-only keys.
- 7 Создать explainable reconcile report.

Обогатить факты справочником

- 1 Primary: строки фактов, например `orders`.
- 2 Source: уникальный справочник, например `parts_catalog`.
- 3 Режим: **Присоединить столбцы по ключу**.
- 4 Убедиться, что source-side duplicate rows равны нулю.
- 5 Запустить `N:1` join.

Troubleshooting (что делать при проблемах)

Сообщение или симптом	Что означает	Что делать
Столбец источника не сопоставлен	Он не попадёт в вертикальный union	Выберите target или сознательно оставьте его
Целевой столбец останется пустым	У источника нет значения для output column	Проверьте mapping; <code>NULL</code> может быть ожидаем
Значения будут преобразованы в <code>NULL</code>	Часть source values не приводится к типу primary	Исправьте тип/данные источника или выберите другой target
Низкое покрытие ключа	Большинство ключей не совпадает	Проверьте выбранные поля, формат и направление primary/source
Совпадения исчезли после добавления пары ключа	Новая часть ключа содержит различия, <code>NULL</code> или несовместимый формат	Посмотрите предупреждение с именем добавленной пары; перенесите обычное проверяемое поле в «Сравнить значение»
<code>different</code> из-за <code>Текст/текст</code>	Включено точное сравнение регистра	Включите «Не учитывать регистр текста»; ключи при этом останутся строгими
<code>only_primary</code> , но строки кажутся похожими	Точного совпадения ключа нет; возможна опечатка	Проверьте примеры unmatched keys. Программа не угадывает кандидата и не показывает вымышленные различия
Есть пустые ключи	В одной из частей ключа <code>NULL</code>	Заполните ключ или исключите такие строки заранее
<code>1:N</code> или <code>N:N</code> , join заблокирован	Источник не уникален по ключу	Уточните ключ, дедуплицируйте или агрегируйте источник
Preset применён частично	Схема одного из новых датасетов изменилась	Исправьте отсутствующие source/target связи; режим, ключи и сравниваемые поля уже восстановлены отдельно
<code>409 Conflict</code> после preview	Конфигурация или версия входа изменилась	Дождитесь нового preview и проверьте его

Сообщение или симптом	Что означает	Что делать
Старый preview исчез после рестарта	Signing key процесса изменился	Запустите preview заново
Нужны все проблемные строки, а не несколько примеров	Preview показывает короткую диагностическую выборку	Создайте результат Эксперт или Reconcile и работайте с полным датасетом

Примеры на Auto Parts

Проверка включения одного списка в другой

Контрольный список **A** содержит детали **P001** и **P002**. Полный каталог **B** содержит **P001**, **P002** и **P003**.

Операция	Результат	Вывод
A - B	0 строк	Все детали из A присутствуют в B
B - A	Полная строка P003	Эта деталь есть только в B

Если названия или цены **P001** различаются, результат **A - B** всё равно будет пустым: Эксперт сравнивает выбранный ключ. Для проверки цены используйте Reconcile.

Сверка обновления цены

В качестве Primary выберите **parts_catalog**, источником - **price_update**. Ключом свяжите **sku** → **sku**, а для сравнения выберите **new_price** → **unit_price**.

Preview покажет:

- **SUS-00001** - ключ и цена совпали;
- **BDY-00001** - цена изменилась с **215.76** на **999.0**;
- **UNKNOWN-999** - ключ есть только в источнике;
- остальные детали каталога - только в основной таблице.

Обогащение каталога справочником брендов

В качестве Primary выберите **parts_catalog**, источником - **brands**, ключом свяжите **brand_id** → **brand_id**. Кардинальность **N:1** означает, что много деталей могут относиться к одному уникальному бренду. Join Columns добавит к каждой детали название, страну и уровень бренда, сохранив число строк каталога.

Диагностика неуникального источника

Если попытаться присоединить **orders** к каталогу деталей, один **part_id** источника может встретиться несколько раз. Preview покажет кардинальность **1:N** и примеры повторяющихся ключей. В зависимости от задачи сделайте **orders** основной таблицей либо сначала агрегируйте заказы до одной строки на деталь.