

Приоритетные правила типов

Безопасный импорт ИНН, КПП, БИН и других идентификаторов

Результат. Значение 0012345678 остаётся строкой с ведущими нулями. Правила принадлежат рабочему пространству, а ручной выбор типа в предпросмотре всегда имеет финальный приоритет.

Правило	Заголовки	Проверка	Результат
ИНН	ИНН, INN	10 или 12 цифр	Текст
КПП	КПП, КРР	9 цифр	Текст
БИН	БИН, VIN	12 цифр	Текст

Обычное определение типа считает последовательность цифр числом. Для идентификаторов это опасно: ведущие нули теряются. DATAŁAD сначала получает обычную рекомендацию, затем проверяет подходящее правило и показывает оператору прозрачную диагностику.

Короткий маршрут

- Датасеты → Импорт → выберите CSV или Excel и рабочее пространство.
- Раскройте «Приоритетные правила типов» и при необходимости измените список.
- Проверьте строку рекомендации под типом столбца и запустите импорт.

Редактор правил рабочего пространства

Настройки одного workspace не влияют на другой. Встроенные правила можно отключить или изменить; пользовательские — добавить и удалить.

Приоритетные правила типов

Правила рабочего пространства защищают идентификаторы ИНН, КПП и БИН от преобразования в число.

+ Добавить правило

Если столбцу подходят несколько правил, применяется правило с большим приоритетом.

ИНН

по умолчанию

#300

Название правила

ИНН

Приоритет

300

Рекомендуемый тип

Текст

Названия столбца (через запятую)

ИНН, INN

Регулярное выражение для названия (необязательно)

Допустимая длина (через запятую)

10, 12

Регулярное выражение для значения (необязательно)

☒ Только цифры

КПП

по умолчанию

#200

БИН

по умолчанию

#100

Вернуть правила по умолчанию

Сохранить правила

Что настраивается

- Приоритет: если подходят несколько правил, побеждает большее число.
- Алиасы и regex заголовка: сопоставление по имени столбца без учёта регистра.
- Длина, «только цифры» и regex значения: проверка непустых значений столбца.
- Рекомендуемый тип: тип, который интерфейс предложит после полного совпадения.

После «Сохранить правила» открытый предпросмотр обновляется автоматически.
Кнопка «Вернуть правила по умолчанию» удаляет только workspace-настройки.

Как читать рекомендацию

Нижe — реальный предпросмотр CSV с ИНН, КПП и БИН. Ведущие нули видны в исходных строках, а тип «Текст» выбран правилом автоматически.

Н заголовок: 1 строка

1

ИНН

КПП

БИН

Компания

2

0012345678

001234567

001234567890

Альфа

3

123456789012

123456789

123456789012

Бета

► Указать точные строки

Проверьте столбцы — строк: 2. Выберите тип для каждого столбца; значения, которые не подходят, станут пустыми.

ИТОГОВЫЕ СТОЛБЦЫ (4)

ИНН

КПП

БИН

Компания

СТОЛБЕЦ	ТИП	ДАННЫЕ	ПРОВЕРКА
ИНН	Текст Правило «ИНН»: совпало значений — 2 → рекомендовано	0% пусто · уникальных: 2 напр. 0012345678 · 123456789012	✓ всё подходит
КПП	Текст Правило «КПП»: совпало значений — 2 → рекомендовано	0% пусто · уникальных: 2 напр. 001234567 · 123456789	✓ всё подходит
БИН	Текст Правило «БИН»: совпало значений — 2 → рекомендовано	0% пусто · уникальных: 2 напр. 001234567890 · 123456789012	✓ всё подходит
Компания	Текст определён	0% пусто · уникальных: 2 напр. Бета · Альфа	✓

Диагностика	Что означает
совпало 2 → рекомендовано	Все непустые значения прошли правило; предложен целевой тип.
есть некорректные значения	Правило не заменяет обычный inference; проверьте данные.
определён	Подходящего приоритетного правила нет; используется обычная рекомендация.

Своё правило: код клиента

Поле	Пример
Название / алиасы	Код клиента / customer_id, client_code
Regex значения	^[A-Z]{2}-[0-9]{8}\$
Тип / приоритет	text / 500

Не сочетайте слишком широкое regex названия с высоким приоритетом. Сначала проверьте диагностику в предпросмотре. При сомнении вручную выберите «Текст».

Производительность и реальные тесты

Почему правила не замедляют импорт заметно

- YARA не используется: правила заранее фильтруются по имени столбца.
- Для столбца применяется одно включённое правило с наибольшим приоритетом.
- Встроенные длина/цифры работают без regex; пользовательский regex компилируется один раз на столбец.
- Значения проверяются в существующем проходе CSV/XLSX; отдельного повторного чтения файла нет.
- Статистика предпросмотра добавлена в существующий агрегирующий запрос DataFusion.

Сценарий	Проверенный результат	Статус
ИНН 0012345678	Parquet хранит строку с ведущим нулём	✓ Rust
ИНН 10 и 12 цифр	text; совпало 2 из 2	✓ Rust
ИНН из 5 цифр	обычный inference без принудительного text	✓ Rust
Изоляция workspace	чужие настройки недоступны	✓ DB
Сохранение UI	PUT содержит правила выбранного workspace	✓ Playwright
Предпросмотр UI	workspace_id передан; выбран text	✓ Playwright

API и хранение

SQLite хранит только разницу относительно системных defaults. Доступ проверяется по владельцу workspace. Ограничения: до 64 правил; regex до 512 байт; выражения валидируются при сохранении.

```
GET /api/v1/workspaces/{id}/import-type-rules
PUT /api/v1/workspaces/{id}/import-type-rules
DELETE /api/v1/workspaces/{id}/import-type-rules
```

Быстрая диагностика

- Правило не показано: проверьте workspace, алиас/regex заголовка и флаг включения.
- Правило найдено, но не применено: хотя бы одно непустое значение не прошло проверку.
- Несколько правил подходят: применяется наибольший приоритет.
- Нужно гарантировать исходный текст: вручную выберите «Текст» в предпросмотре.