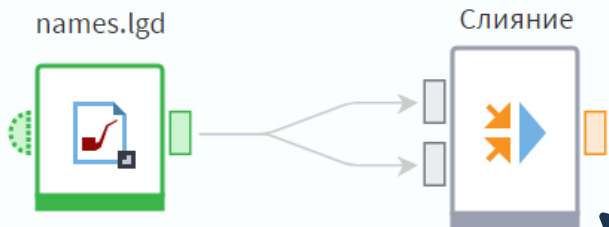


The background of the slide features a person's hands typing on a laptop keyboard, overlaid with a semi-transparent blue layer. Several glowing blue data packets, each containing binary code and a padlock icon, are scattered across the scene, connected by faint lines, suggesting data flow or security. The overall aesthetic is high-tech and digital.

# МЕТОДЫ ОЧИСТКИ ДАННЫХ: АНАЛИЗ СТРОК

#	ab ID	ab ФИО	12 Год рождения
1	1	Плотников Константин Леонидович	1985
2	2	Гасимова Юлия Ивановна	1985
3	3	Запрудина Любовь Ивановна	1992
4	4	Проخورова Ольга Германовна	1986
5	5	Плотникова Елена Александровна	1975
6	6	Гасимова Юлия Ивагновна	1985
7	7	Запрудина Любовь Ивановна	1992
8	8	Котова Екатерина Владимировна	1997
		сандрович	1979
		ич	1990
		сандровна	1975
		мировна	1997
		вич	1990
		ндрович	1979
		мировна	1984
		довна	1998
		вна	
		на	1991

Создадим новый модуль в нашем пакете и импортируем файл **names.lgd**, в котором содержится несколько десятков записей — ФИО клиентов магазина и их год рождения. Нам нужно понять, есть ли в этом списке дубликаты. Для их обнаружения есть разные пути, но мы воспользуемся расстояниями Левенштейна и Дамерау-Левенштейна.



Нам необходимо сравнить каждое ФИО в нашем списке с каждым, поэтому для начала проведем операцию полного соединения с помощью узла слияния.

Тип операции

Полное соединение

Фильтрация

Столбцы основного набора данных

ab ID

ab ФИО

12 Год рождения

Фильтрация

Столбцы присоединяемого набора данных

ab ID

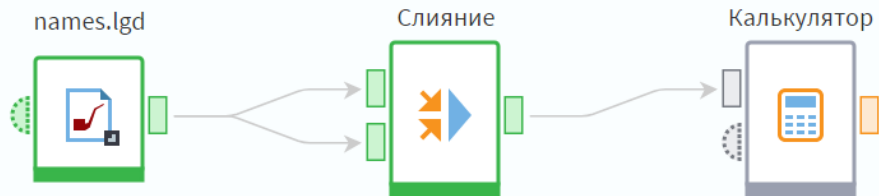
ab ФИО

12 Год рождения

В настройках узла не требуется добавлять связи между полями.

#	ab ID	ab ФИО	12 Год рождения	ab ID	ab ФИО	12 Год рождения
1	1	Плотников Константин Леонидович	1985	1	Плотников Константин Леонидович	1985
2	1	Плотников Константин Леонидович	1985	2	Гасимова Юлия Ивановна	1985
3	1	Плотников Константин Леонидович	1985	3	Запрудина Любовь Ивановна	1992
4	1	Плотников Константин Леонидович	1985	4	Прохорова Ольга Германовна	1986
5	1	Плотников Константин Леонидович	1985	5	Плотникова Елена Александровна	1975
6	1	Плотников Константин Леонидович	1985	6	Гасимова Юлия Ивагновна	1985
7	1	Плотников Константин Леонидович	1985	7	Запрудина Любовь Ивановна	1992
8	1	Плотников Константин Леонидович	1985	8	Котова Екатерина Владимировна	1997
9	1	Плотников Константин Леонидович	1985	9	Бородин Николай Александрович	1979
10	1	Плотников Константин Леонидович	1985	10	Запрудина Мария Леонидовна	1990
11	1	Плотников Константин Леонидович	1985	11	Гасимова Юлия Ивановна	1975
12	1	Плотников Константин Леонидович	1985	12	Запрудина Любовь Ивановна	1997
13	1	Плотников Константин Леонидович	1985	13	Прохорова Ольга Германовна	1990
14	1	Плотников Константин Леонидович	1985	14	Плотников Константин Леонидович	1979
15	1	Плотников Константин Леонидович	1985	15	Гасимова Юлия Ивановна	1984
16	1	Плотников Константин Леонидович	1985	16	Запрудина Мария Леонидовна	1998
17	1	Плотников Константин Леонидович	1985	17	Минина Оксана Николаевна	
18	1	Плотников Константин Леонидович	1985	18	Титова Елена Михайловна	1991
19	1	Плотников Константин Леонидович	1985	19	Броодин Николай Александрович	1979
20	1	Плотников Константин Леонидович	1985	20	Ягодин Виктор Александрович	1987
1 225	1	Плотников Константин Леонидович	1985	21	Калинина Светлана Валентиновна	1982

На выходе мы получим нужный для сравнения значений набор.



Вспользуемся узлом **Калькулятор**. Мы уже знаем, что в нем есть нужные нам функции.

Имя	Метка
12 Lev	Расстояние Левенштейна
12 DevLev	Расстояние Дамерау-Левен...

Предпросмотр... | AND OR NOT XOR | = <> < > <= >= | 9.0 ≡

```
LevDist(FullName,FullName_1)
```

Создадим два целочисленных поля для расчета расстояний, в каждом воспользуемся соответствующей функцией, синтаксис функций идентичен.

Поля/Переменные

Имя	Метка
Поля	
ab ID	ID
ab FullName	ФИО
12 BirthYear	Год рождения
ab ID_1	ID
ab FullName_1	ФИО
12 BirthYear_1	Год рождения

- Список фун
- 9.0 Abs (Арг)
  - 9.0 AbsErr (Аргумент1, Аргумент2)
  - 31 AddDay (Дата, Количество)
  - 31 AddMonth (Дата, Количество)
  - 31 AddQuarter (Дата, Количество)
  - 31 AddWeek (Дата, Количество)
  - 31 AddYear (Дата, Количество)
  - 9.0 AMGD (Стоимость, Остаточная\_стоимость, Время...
  - 9.0 ArcCos (Значение)
  - 9.0 ArcSin (Значение)

Выражения

*f(x)* | ^ v

Имя	Метка
12 Lev	Расстояние Левенштейна
12 DevLev	Расстояние Дамерау-Левен...

Предпросмотр... | AND OR NOT XOR | = <> < > <= >= | 9.0

DamLevDist(FullName, FullName\_1)

Поля/Переменные

Фильтрация

Имя	Метка
Поля	
ab ID	ID
ab FullName	ФИО
12 BirthYear	Год рождения
ab ID_1	ID
ab FullName_1	ФИО
12 BirthYear_1	Год рождения

Список функций

леве

Категории

12 DamLevDist (Строка1, Строка2)
12 LevDist (Строка1, Строка2)

**DamLevDist(Строка1, Строка2)**

Функция возвращает значение расстояния Дамерау-Левенштейна для строк Строка1, Строка2. Расстояние Дамерау-Левенштейна также называют расстоянием редактирования с учетом перестановок, которое является мерой схожести двух строк. Результат - это минимальное количество операций удалений, вставки.



Выражения

*fix* |

Предпросмотр... | AND OR NOT XOR | = <> < > <= >= | 9.0

Имя	Метка
12 Lev	Расстояние Левенштейна
12 DevLev	Расстояние Дамерау-Левен...
0/1 SameID	Совпадает ID

```
ID=ID_1
```

Дополнительно создадим два логических поля. В первом проверяем, совпадает ли в записях ID, тогда значение **true** будет означать, что запись сравнивается сама с собой.

Поля/Переменные

Фильтрация

Список функций

леве

Категории

Имя	Метка
Поля	
ab ID	ID
ab FullName	ФИО
12 BirthYear	Год рождения
ab ID_1	ID
ab FullName_1	ФИО

12 DamLevDist (Строка1, Строка2)
12 LevDist (Строка1, Строка2)

**DamLevDist(Строка1, Строка2)**  
Функция возвращает значение расстояния Дамерау-Левенштейна для строк Строка1, Строка2. Расстояние Дамерау-Левенштейна также называют расстоянием редактирования с учетом перестановок, которое является мерой похожести двух

Выражения *fix* |

Имя	Метка
12 Lev	Расстояние Левенштейна
12 DevLev	Расстояние Дамерау-Левен...
0/1 SameID	Совпадает ID
0/1 SameYear	Совпадает год

Предпросмотр... | AND OR NOT XOR | = <> < > <= >= | 9.0

`BirthYear=BirthYear_1`

Во втором проверяем, совпадает ли год рождения.

Поля/Переменные

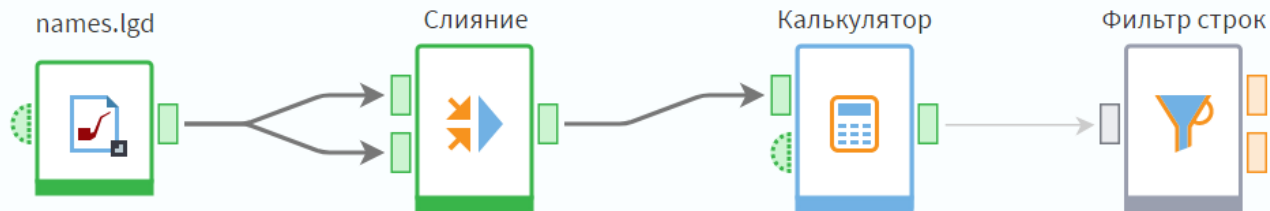
Имя	Метка
ab ID	ID
ab FullName	ФИО
12 BirthYear	Год рождения
ab ID_1	ID
ab FullName_1	ФИО
12 BirthYear_1	Год рождения

Список функций

- 12 DamLevDist (Строка1, Строка2)
- 12 LevDist (Строка1, Строка2)

Категории

**DamLevDist(Строка1, Строка2)**  
Функция возвращает значение расстояния Дамерау-Левенштейна для строк Строка1, Строка2. Расстояние Дамерау-Левенштейна также называют расстоянием редактирования с учетом перестановок, которое является мерой похожести двух



Сравнивать запись саму с собой нам неинтересно, поэтому отфильтруем только те записи, где в поле **Совпадает ID** значение **false**.

Состояние входа

Вход активирован

[Активировано](#)

0/1 Совпадает ID = ЛОЖЬ

×

+

Соответствуют условию    Не соответствуют условию

#	12 Расстояние Левенштейна	12 Расстояние Дамерау-Левенштейна	0/1 Совпадает ID	0/1 Совпадает год	ab ID	ab ФИО	12 Год рождения
1	23	23	false	true	1	Плотников Константин Леонидович	1985
2	26	26	false	false	1	Плотников Константин Леонидович	1985
3	22	22	false	false	1	Плотников Константин Леонидович	1985
4	18	18	false	false	1	Плотников Константин Леонидович	1985
5	24	24	false	true	1	Плотников Константин Леонидович	1985
6	26	26	false	false	1	Плотников Константин Леонидович	1985
7	23	23	false	false	1	Плотников Константин Леонидович	1985
8	23	23	false	false	1	Плотников Константин Леонидович	1985
9	22	22	false	false	1	Плотников Константин Леонидович	1985
10	18	18	false	false	1	Плотников Константин Леонидович	1985
11	22	22	false	false	1	Плотников Константин Леонидович	1985
12	22	22	false	false	1	Плотников Константин Леонидович	1985
13	20	20	false	false	1	Плотников Константин Леонидович	1985
14						ин Леонидович	1985
15						ин Леонидович	1985
16						ин Леонидович	1985
17						ин Леонидович	1985
18						ин Леонидович	1985
19						ин Леонидович	1985
20	22	22	false	false	1	Плотников Константин Леонидович	1985
1 190							

Получили набор из 1190 записей, как показано на слайде. Для большинства записей значения расстояний большие. Они нас не интересуют, так как не могут являться дубликатами.

names.lgd



Слияние



Калькулятор



Фильтр строк



Фильтр строк



Поэтому воспользуемся еще одним фильтром и отберем записи, где одно из расстояний меньше или равно 2.

Состояние входа

Вход активирован

Активировано

12 Расстояние Левенштейна  $\leq 2$  ×

ИЛИ

12 Расстояние Дамерау-Левенштейна  $\leq 2$  ×

+

Для этого настроим условие, как  
показано на слайде.

Удалить все условия

Соответствуют условию    Не соответствуют условию

#	12 Расстояние Левенштейна	12 Расстояние Дамерау-Левенштейна	0/1 Совпадает ID	0/1 Совпадает год	ab ID	ab ФИО	12 Год рождения
1	1	1	false	true	2	Гасимова Юлия Ивановна	1 985
2	0	0	false	true	3	Запрудина Любовь Ивановна	1 992
3	0	0	false	true	5	Плотникова Елена Александровна	1 975
4	1	1	false	true	6	Гасимова Юлия Ивагновна	1 985
5	0	0	false	true	7	Запрудина Любовь Ивановна	1 992
6	1	1	false	true	8	Котова Екатерина Владимировна	1 997
7	1	1	false	true	8	Котова Екатерина Владимировна	1 997
8	1	1	false	true	9	Котлов Николай Александрович	1 979
9	1	1	false	true	9	Котлов Максим Андреевич	1 990
10	1	1	false	true	10	Плотникова Елена Александровна	1 975
11	1	1	false	true	10	Котова Екатерина Владимировна	1 997
12	1	1	false	true	10	Котова Екатерина Владимировна	1 997
13	1	1	false	true	11	Котова Екатерина Владимировна	1 984
14	1	1	false	true	11	Котова Екатерина Владимировна	1 984
15	1	1	false	true	12	Котлов Николай Александрович	1 979
16	1	1	false	true	12	Котлов Максим Андреевич	1 990
17	1	1	false	true	13	Котова Вероника Николаевна	1 975
18	1	1	false	true	13	Котова Вероника Николаевна	1 975

На выходе остается 18 записей. Но не обязательно все сравнения в них являются дубликатами. ФИО часто бывают похожи или даже полностью идентичны, поэтому одного расстояния не достаточно для однозначного решения. Самый надежный вариант, если в данных также присутствует номер. паспорта, можно также ориентироваться на совпадение адреса. В нашем случае у нас есть только год рождения, поэтому будем считать, что если расстояние небольшое, и год совпал, то исходные записи являются дублиями

Соответствуют условию    Не соответствуют условию

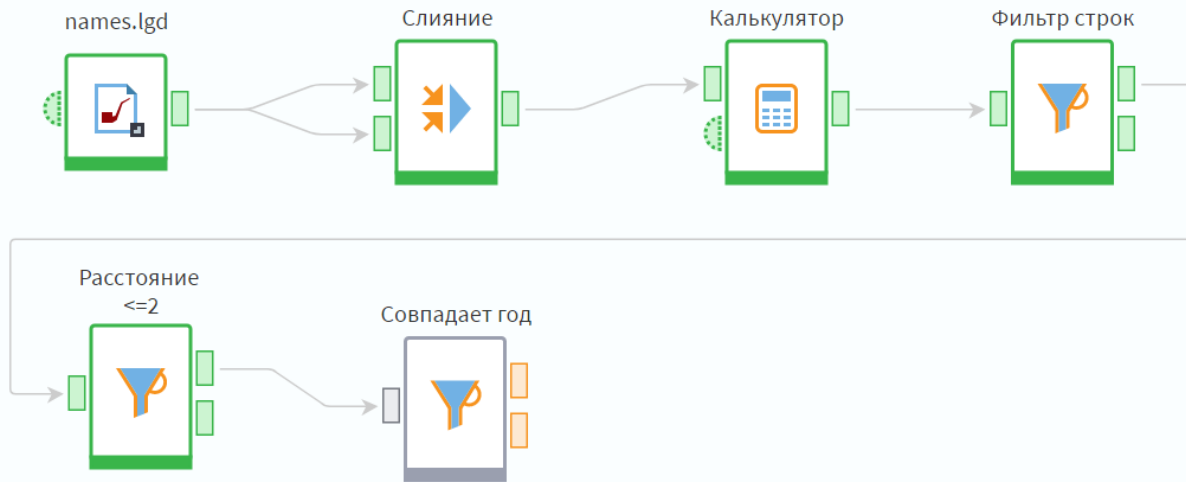
#	12 Расстояние Левенштейна	12 Расстояние Дамерау-Левенштейна	0/1 Совпадает ID	0/1 Совпадает год	ab ID	ab ФИО	12 Год рождения
1	1	1	false	true	2	Гасимова Юлия Ивановна	1 985
2	0	0	false	true	3	Запрудина Любовь Ивановна	1 992
3	0	0	false	true	5	Плотникова Елена Александровна	1 975
4	1	1	false	true	6	Гасимова Юлия Ивагновна	1 985
5	0	0	false	true	7	Запрудина Любовь Ивановна	1 992
6	1	1	false	true	8	Котова Екатерина Владимировна	1 997
7	1		false	false	8	Котова Екатерина Владимировна	1 997
8	2		false	true	9	Бородин Николай Александрович	1 979
9							1 990
10							1 975
11							1 997
12							1 997
13							1 984
14							1 984
15							1 979
16	1	1	false	true	24	Руднев Максим Адреевич	1 990
17	1	1	false	true	33	Антипова Вероника Николавна	1 975
18	1	1	false	true	35	Антипова Вероника Николаевна	1 975

Например, для первых шести записей так и есть, причем как видно на слайде, записи повторяются, т.к. мы сравнивали каждую с каждой. По сути строки 1 и 4, 2 и 5 идентичны, и нужно выбрать, ФИО с каким 10 мы оставим в данных, а с каким — уберем.



Соответствуют условию		Не соответствуют условию						
#	12 Расстояние Левенштейна	12 Расстояние Дамерау-Левенштейна	0/1 Совпадает ID	0/1 Совпадает год	ab ID	ab ФИО	12 Год рождения	
1	1	1	false	true	2	Гасимова Юлия Ивановна	1 985	
2	0	0	false	true	3	Запрудина Любовь Ивановна	1 992	
3	0	0	false	true	5	Плотникова Елена Александровна	1 975	
4	1	1	false	true	6	Гасимова Юлия Ивагновна	1 985	
5	0	0	false	true	7	Запрудина Любовь Ивановна	1 992	
6	1	1	false	true	8	Котова Екатерина Владимировна	1 997	
7	1	1	false	false	8	Котова Екатерина Владимировна	1 997	
8	1	1	false	true	9	Бородин Николай Александрович	1 979	
9	1	1	false	true	10	Бородин Николай Александрович	1 990	
10	1	1	false	true	11	Бородин Николай Александрович	1 975	
11	1	1	false	true	12	Бородин Николай Александрович	1 997	
12	1	1	false	true	13	Бородин Николай Александрович	1 997	
13	1	1	false	true	14	Бородин Николай Александрович	1 984	
14	1	1	false	true	15	Бородин Николай Александрович	1 984	
15	1	1	false	true	16	Бородин Николай Александрович	1 979	
16	1	1	false	true	24	Гуднев Максим Андреевич	1 990	
17	1	1	false	true	33	Антипова Вероника Николаевна	1 975	
18	1	1	false	true	35	Антипова Вероника Николаевна	1 975	

А вот седьмая строка вряд ли является дублем: **Котова** и **Китова** — скорее всего не опечатка, а действительно разные фамилии, тем более, что год рождения в данном случае отличается.



Оставим только те записи, в которых совпадает год, и отсортируем поля: первый уровень сортировки – поле **FullName** по возрастанию, второй уровень – поле **FullName\_1** по возрастанию.

Состояние входа


Не активировано

[Активировать](#)

0/1 Совпадает год = ИСТИНА ×

+

Оставим только те записи, в которых совпадает год.

 Удалить все условия

Фильтрация

### Доступные поля

- 12 Расстояние Левенштейна
- 12 Расстояние Дамерау-Левенштейна
- 0/1 Совпадает ID
- 0/1 Совпадает год
- ab ID
- 12 Год рождения
- ab ID
- 12 Год рождения

Переместить вверх    Переместить вниз

### Поля сортировки

Поля сортировки	Порядок	Регистр	
ab ФИО		<input checked="" type="checkbox"/>	
ab ФИО		<input checked="" type="checkbox"/>	

Отсортируем поля: первый уровень сортировки – поле **FullName** по возрастанию, второй уровень – поле **FullName\_1** по возрастанию.

Выходной набор данных

#	12 Расстояние Левенштейна	12 Расстояние Дамерау-Левенштейна	0/1 Совпадает ID	0/1 Совпадает год	ab ID	ab ФИО	12 Год рождения
1	1	1	false	true	33	Антипова Вероника Николавна	1975
2	1	1	false	true	35	Антипова Вероника Николаевна	1975
3	2	1	false	true	9	Бородин Николай Александрович	1979
4	2	1	false	true	19	Бродин Николай Александрович	1979
5	1		false	true	6	Гасимова Юлия Ивагновна	1985
6	1		false	true	2	Гасимова Юлия Ивановна	1985
7	0			true	3	Запрудина Любовь Ивановна	1992
8	0						1992
9	1						1997
10	1						1997
11	0						1975
12	0						1975
13	1						1990
14	1						1990

Теперь у нас остались только дублирующиеся записи, легко заметить, что в некоторых присутствуют опечатки. Когда таких записей немного, как в нашем случае, выбрать, какой вариант оставить, а какой — исключить, можно вручную. В промышленных масштабах, когда записей десятки и сотни тысяч, часто используют справочники имен, чтобы проверить корректность написания.