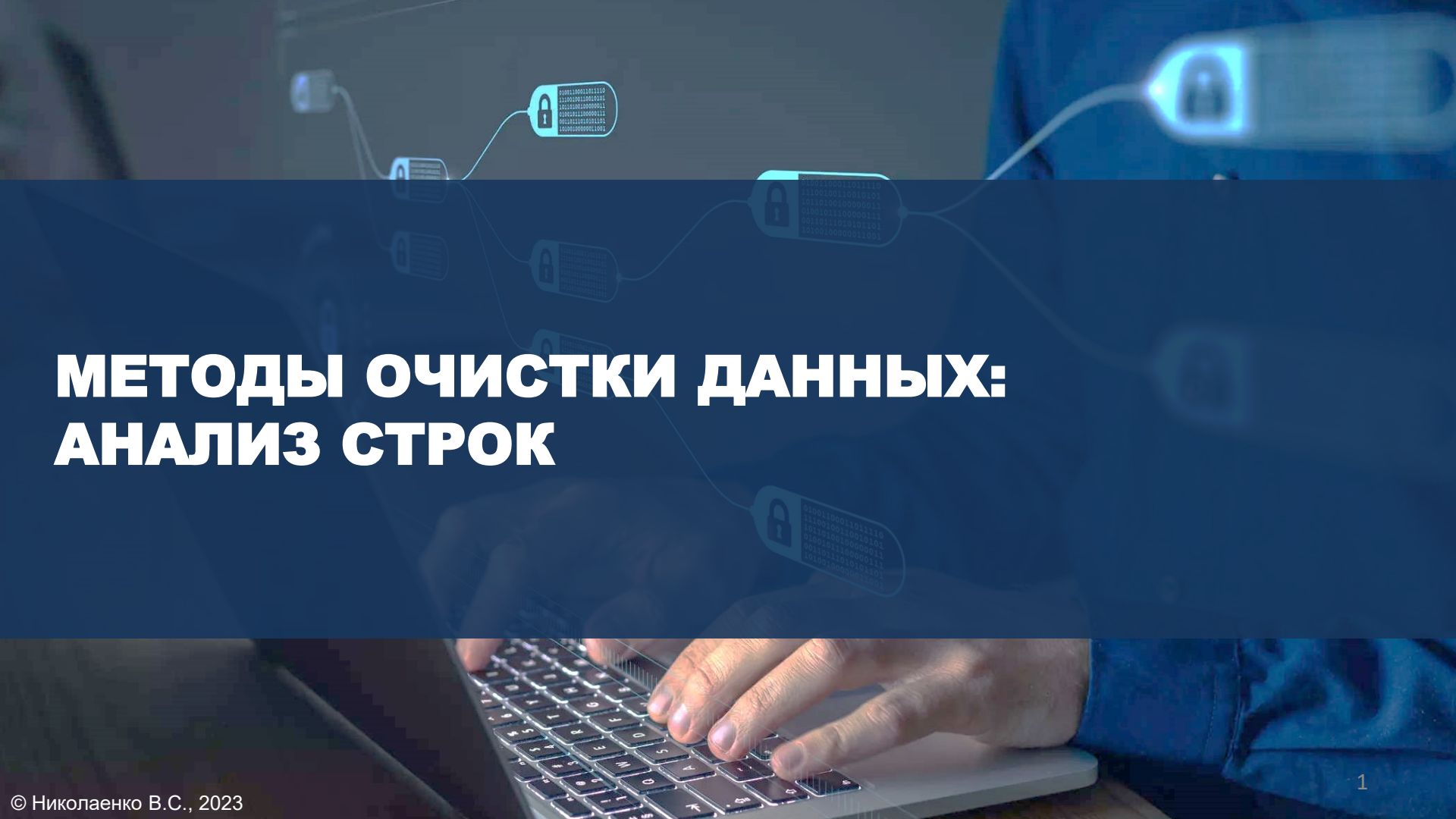
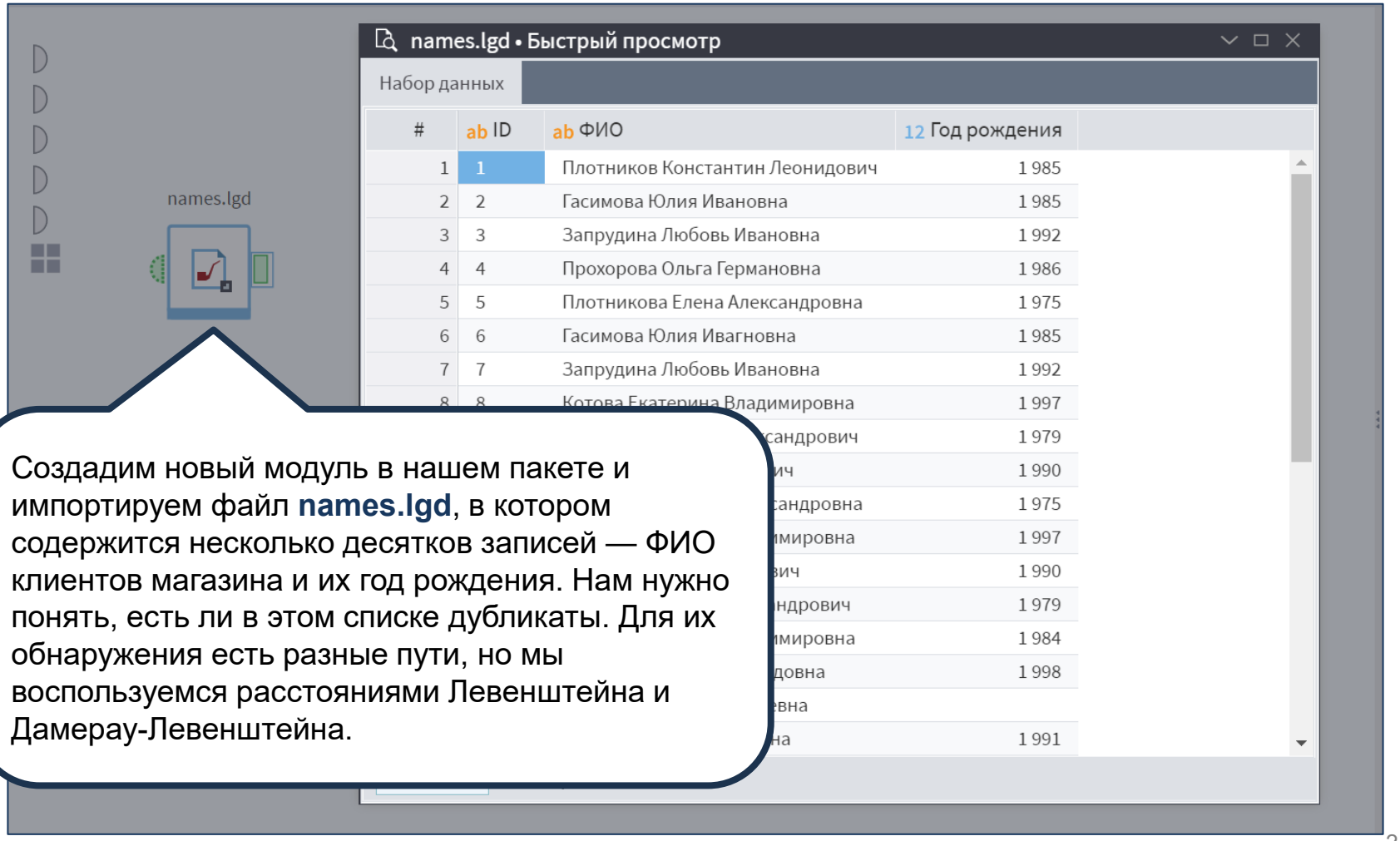
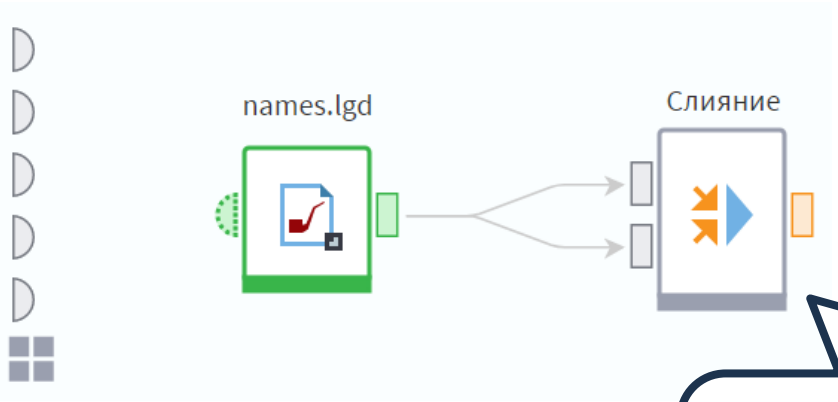


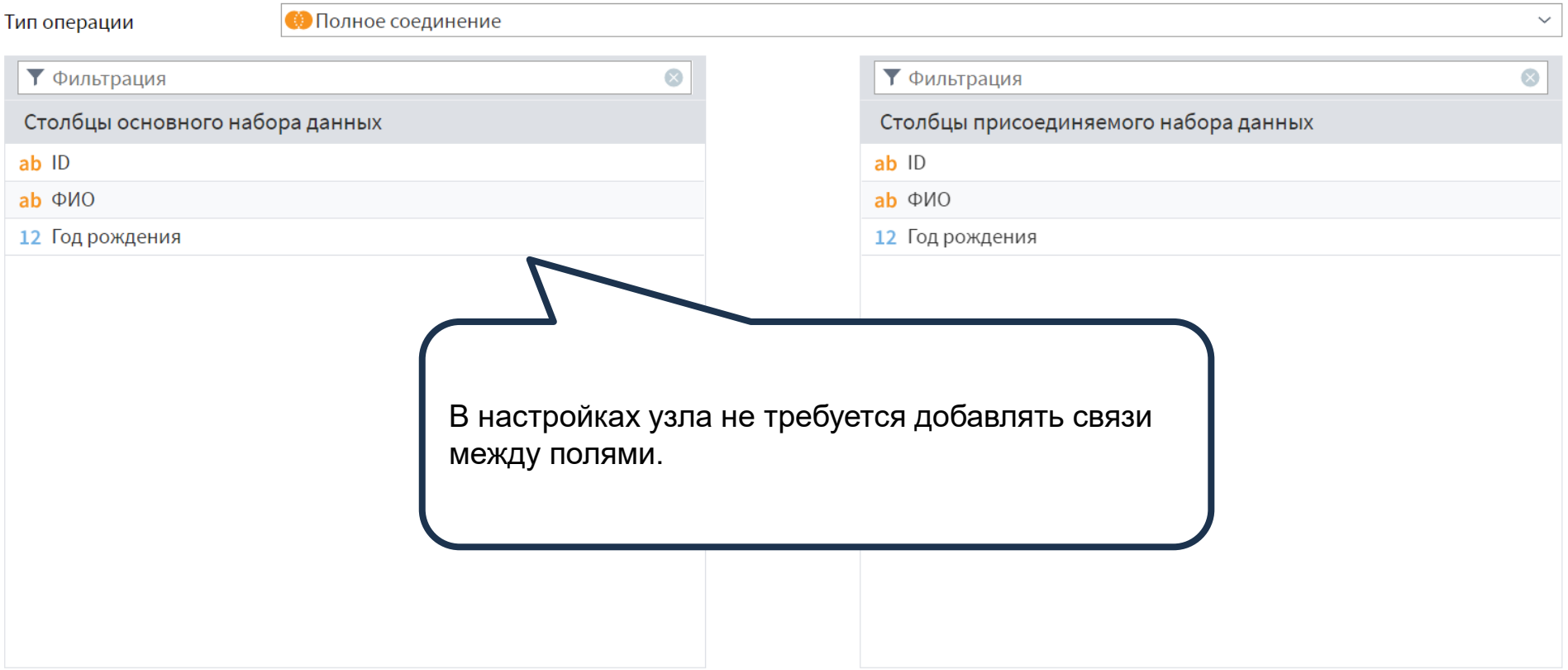
МЕТОДЫ ОЧИСТКИ ДАННЫХ: АНАЛИЗ СТРОК



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.



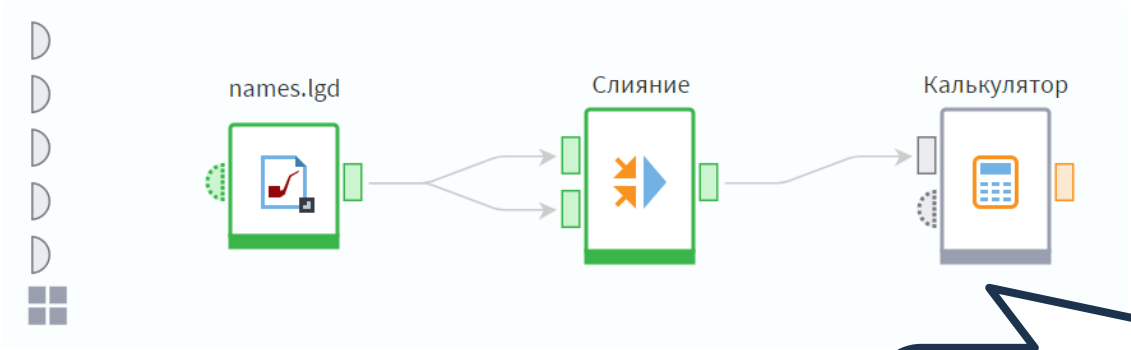
Нам необходимо сравнить каждое ФИО в нашем списке с каждым, поэтому для начала проведем операцию полного соединения с помощью узла слияния.



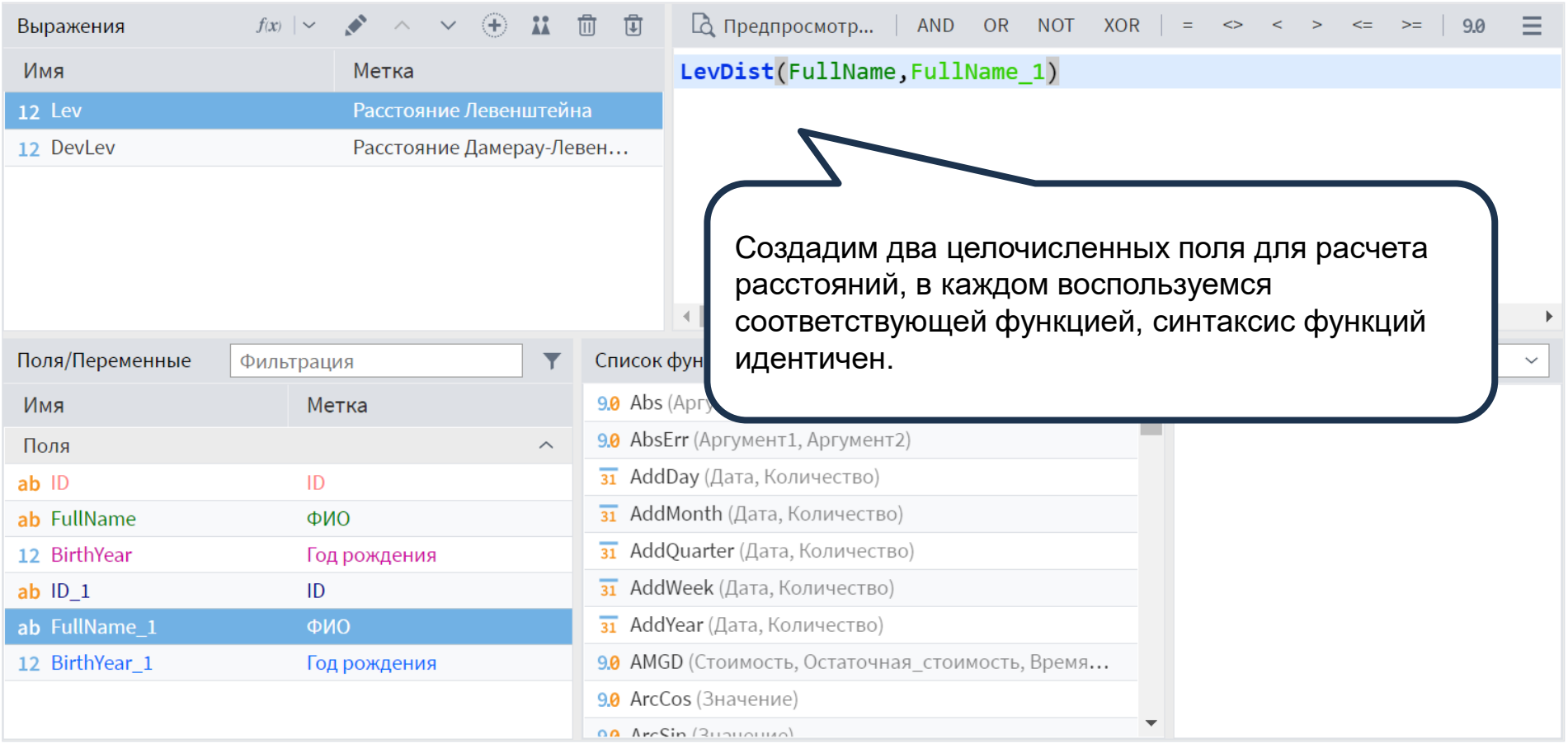
Выходной набор данных

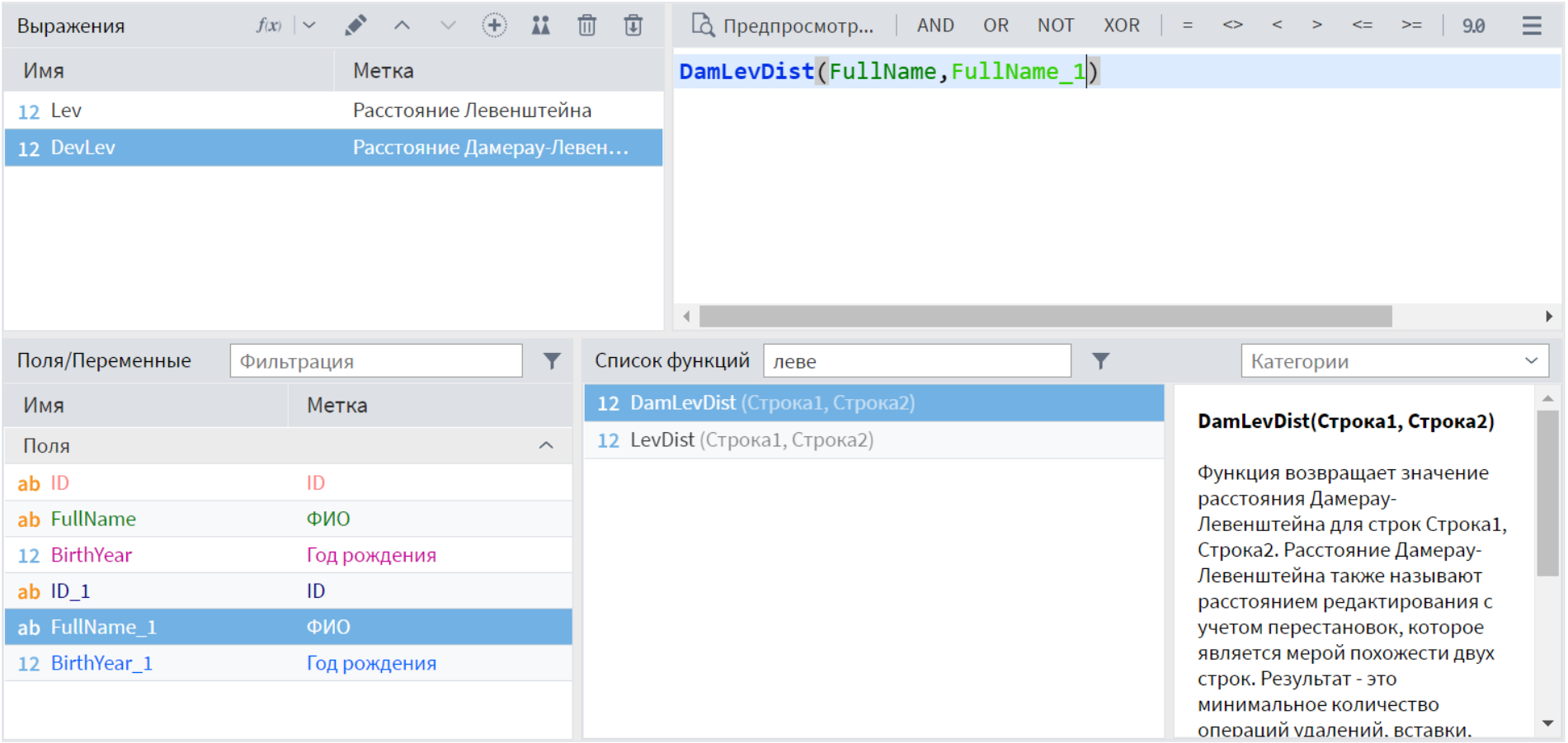
#	ab ID	ab ФИО	12 Год рождения	ab ID	ab ФИО	12 Год рождения
1	1	Плотников Константин Леонидович	1 985	1	Плотников Константин Леонидович	1 985
2	1	Плотников Константин Леонидович	1 985	2	Гасимова Юлия Ивановна	1 985
3	1	Плотников Константин Леонидович	1 985	3	Запрудина Любовь Ивановна	1 992
4	1	Плотников Константин Леонидович	1 985	4	Прохорова Ольга Германовна	1 986
5	1	Плотников Константин Леонидович	1 985	5	Плотникова Елена Александровна	1 975
6	1	Плотников Константин Леонидович	1 985	6	Гасимова Юлия Ивагновна	1 985
7	1	Плотников Константин Леонидович	1 985	7	Запрудина Любовь Ивановна	1 992
8	1	Плотников Константин Леонидович	1 985	8	Котова Екатерина Владимировна	1 997
9	1	Плотников Константин Ле	1 985	9	Бородин Николай Александрович	1 979
10	1	Плотников Кон				1 990
11	1	Плотников Ко			овна	1 975
12	1	Плотников Ко			вна	1 997
13	1	Плотников Ко				1 990
14	1	Плотников Ко			вич	1 979
15	1	Плотников Ко			вна	1 984
16	1	Плотников Константин Леонидович				1 998
17	1	Плотников Константин Леонидович	1 985	17	Минина Оксана Николаевна	
18	1	Плотников Константин Леонидович	1 985	18	Титова Елена Михайловна	1 991
19	1	Плотников Константин Леонидович	1 985	19	Броодин Николай Александрович	1 979
20	1	Плотников Константин Леонидович	1 985	20	Ягодин Виктор Александрович	1 987
1 225	1	Плотников Константин Леонидович	1 985	21	Калинина Светлана Валентиновна	1 982

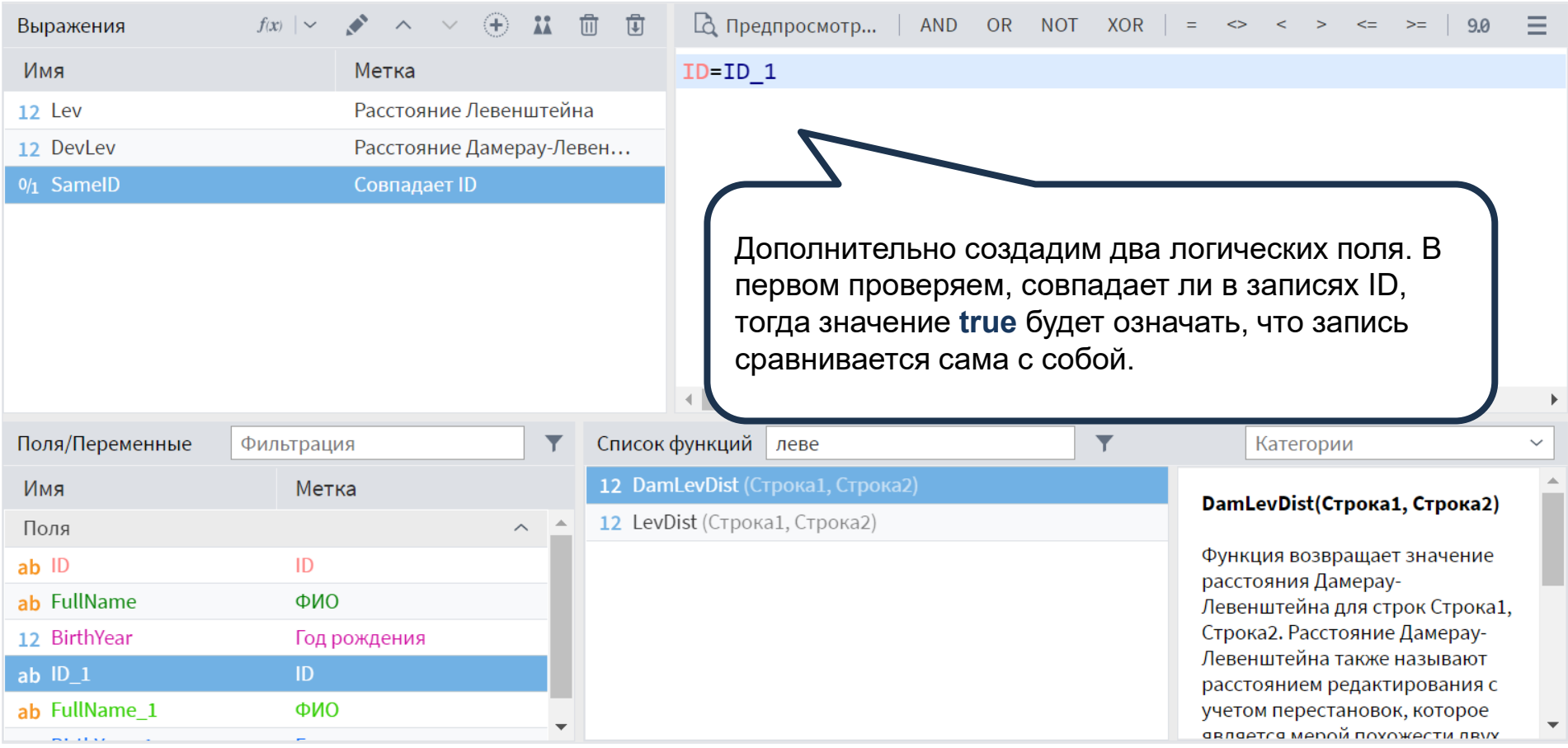
На выходе мы получим нужный для сравнения значений набор.

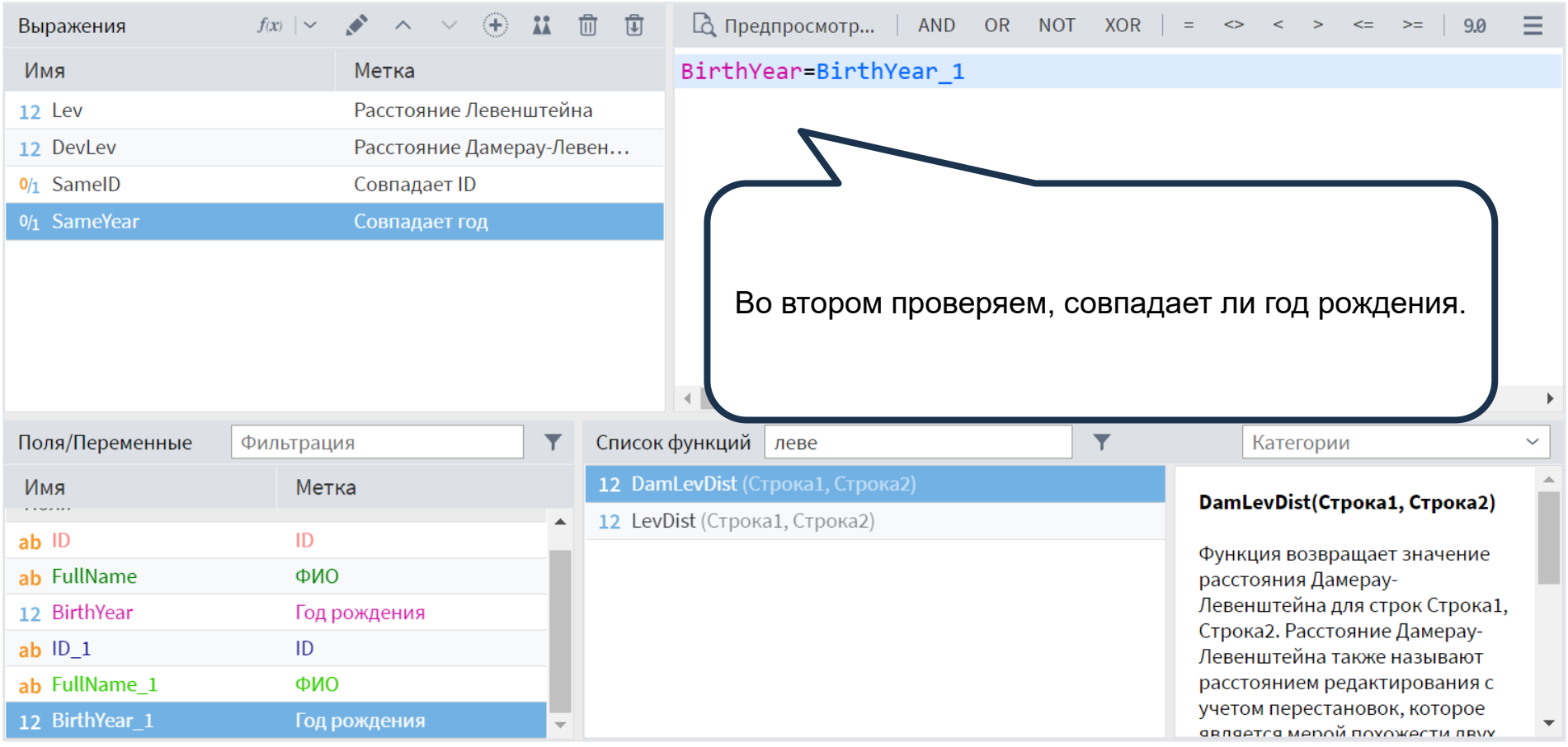


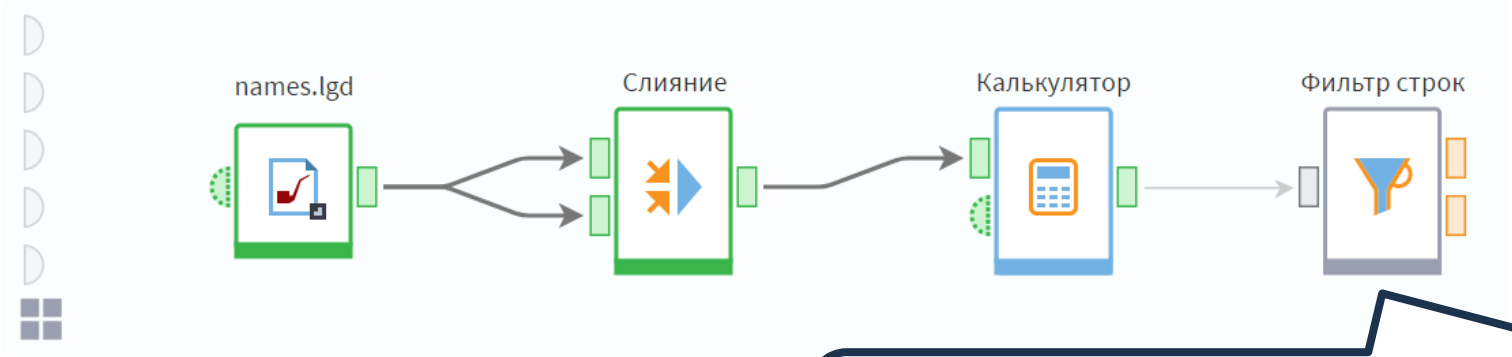
Воспользуемся узлом **Калькулятор**. Мы уже знаем, что в нем есть нужные нам функции.









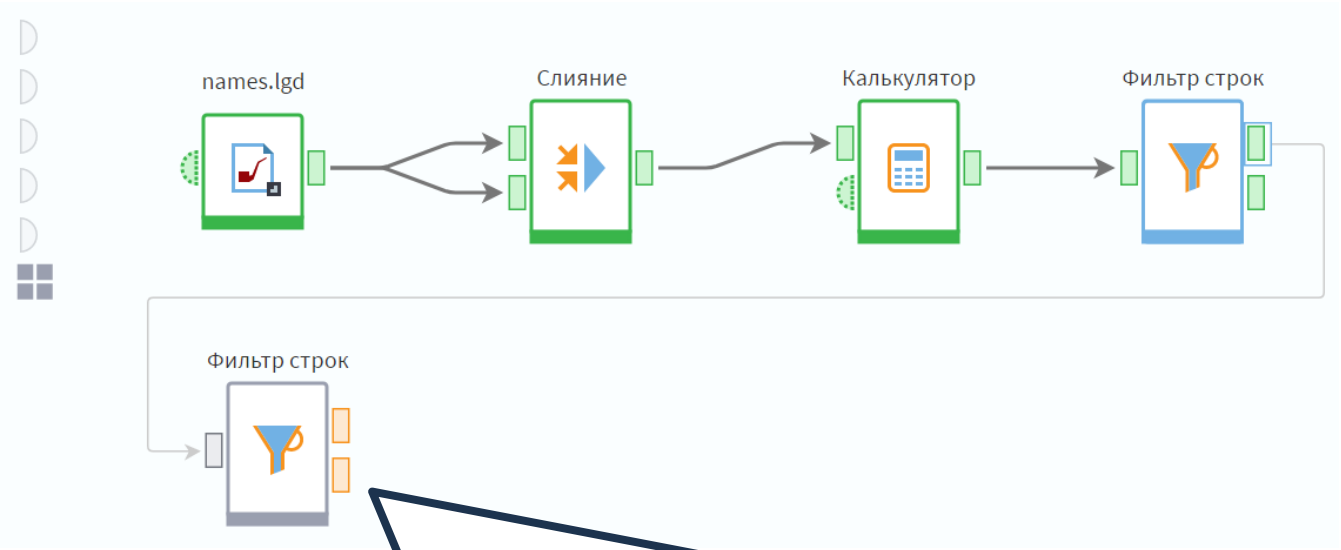


Сравнивать запись саму с собой нам неинтересно, поэтому отфильтруем только те записи, где в поле **Совпадает ID** значение **false**.

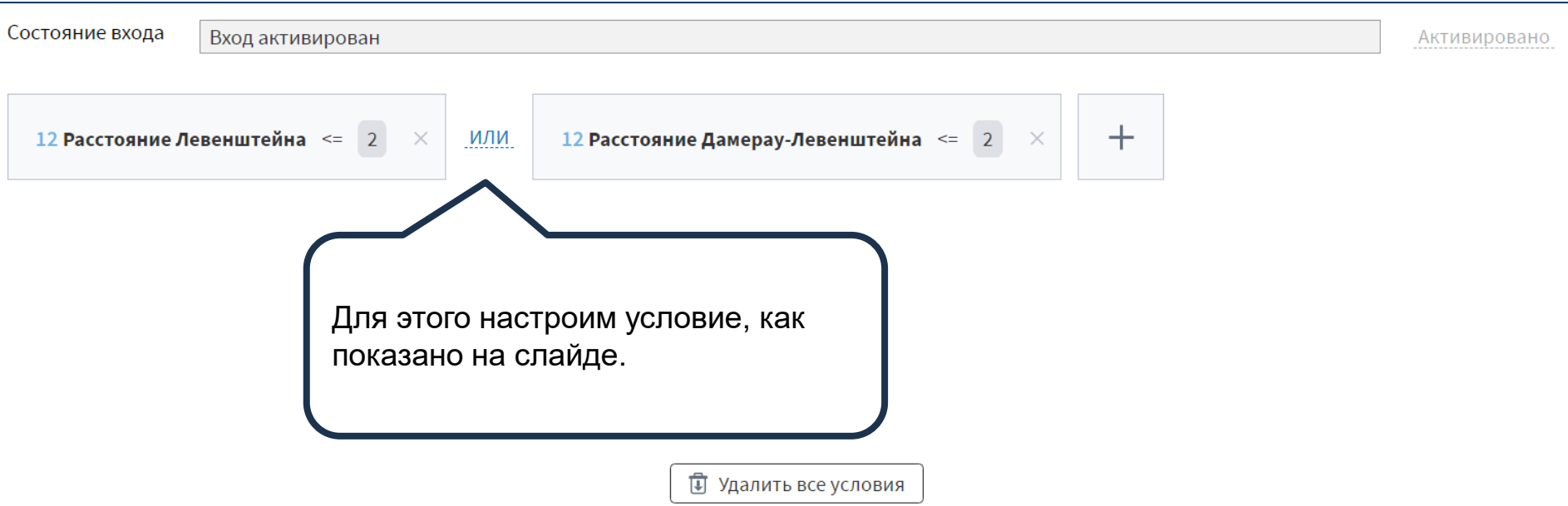
Соответствуют условию | Не соответствуют условию

#	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	1 985
2	26	26	false	false	1	Плотников Константин Леонидович	1 985
3	22	22	false	false	1	Плотников Константин Леонидович	1 985
4	18	18	false	false	1	Плотников Константин Леонидович	1 985
5	24	24	false	true	1	Плотников Константин Леонидович	1 985
6	26	26	false	false	1	Плотников Константин Леонидович	1 985
7	23	23	false	false	1	Плотников Константин Леонидович	1 985
8	23	23	false	false	1	Плотников Константин Леонидович	1 985
9	22	22	false	false	1	Плотников Константин Леонидович	1 985
10	18	18	false	false	1	Плотников Константин Леонидович	1 985
11	22	22	false	false	1	Плотников Константин Леонидович	1 985
12	22	22	false	false	1	Плотников Константин Леонидович	1 985
13	20	20	false	false	1	Плотников Константин Леонидович	1 985
14						ин Леонидович	1 985
15						ин Леонидович	1 985
16						ин Леонидович	1 985
17						ин Леонидович	1 985
18						ин Леонидович	1 985
19						ин Леонидович	1 985
20	22	22	false	false	1	Плотников Константин Леонидович	1 985
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.



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



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						Котова Екатерина Владимировна	1 997
8						дин Николай Александрович	1 979
9						в Максим Андреевич	1 990
10						икова Елена Александровна	1 975
11						а Екатерина Влодимировна	1 997
12						а Екатерина Влодимировна	1 997
13						а Екатерина Владимировна	1 984
14						а Екатерина Владимировна	1 984
15						ин Николай Александрович	1 979
16						в Максим Адреевич	1 990
17						ова Вероника Николавна	1 975
18						ова 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 дублями

Соответствуют условию | Не соответствуют условию

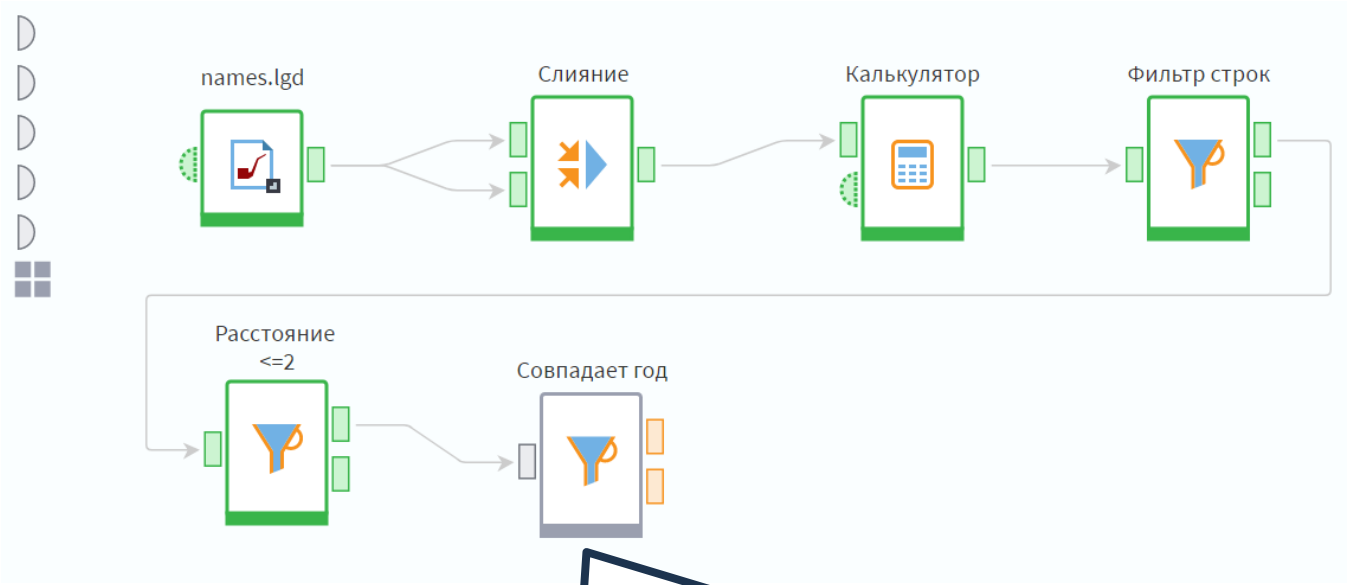
#	12 Расстояние Левенштейна	12 Расстояние Дамерау-Левенштейна	0/1 Совпадает ID	0/1 Совпадает год	ab ID	ab ФИО	12 Год рождения
1	1	1	false	true	2	Гасимова Юлия Ивановна	1 985
2	0	0	false	true	3	Запрудина Любовь Ивановна	1 992
3	0	0	false	true	5	Плотникова Елена Александровна	1 975
4	1	1	false	true	6	Гасимова 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.

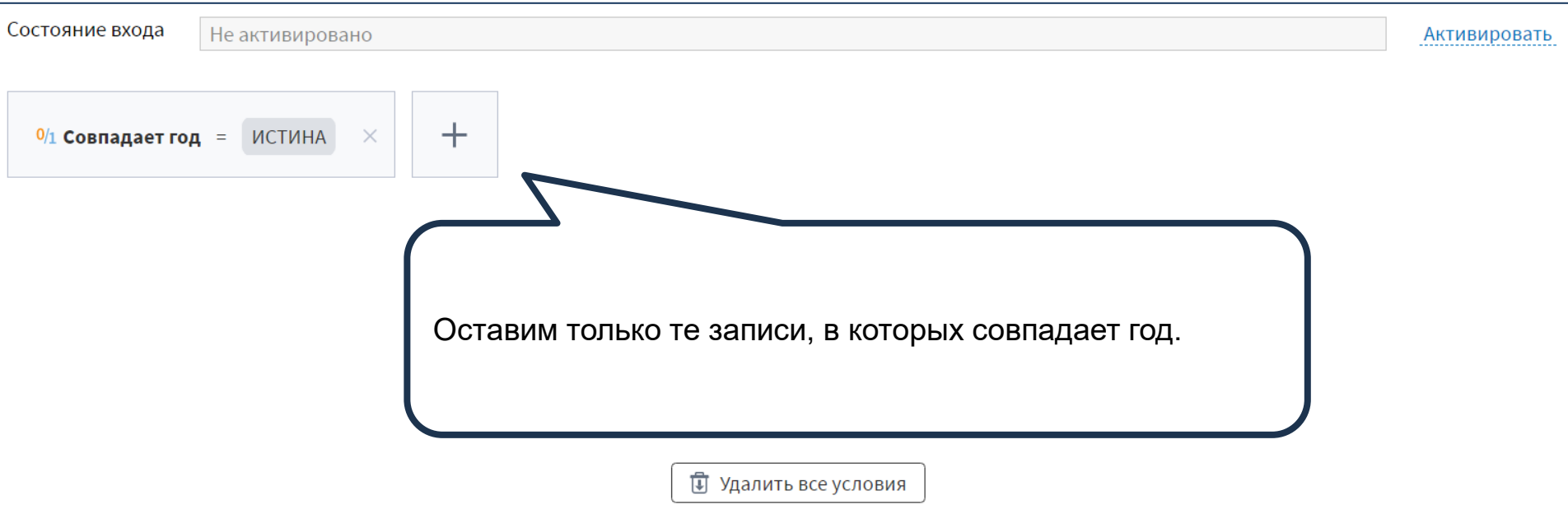
Соответствуют условию | Не соответствуют условию

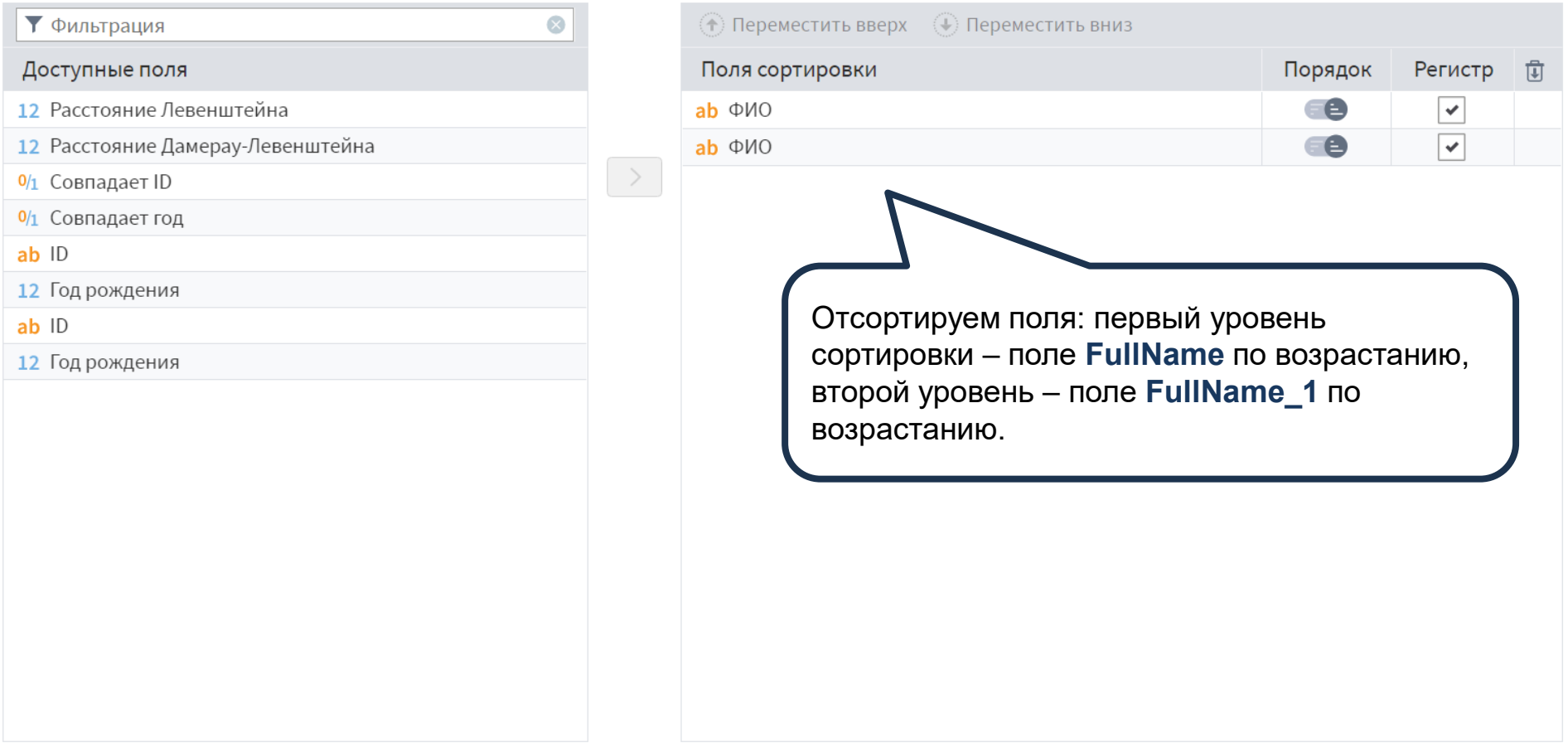
#	12 Расстояние Левенштейна	12 Расстояние Дамерау-Левенштейна	0/1 Совпадает ID	0/1 Совпадает год	ab ID	ab ФИО	12 Год рождения
1	1	1	false	true	2	Гасимова Юлия Ивановна	1 985
2	0	0	false	true	3	Запрудина Любовь Ивановна	1 992
3	0	0	false	true	5	Плотникова Елена Александровна	1 975
4	1	1	false	true	6	Гасимова 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	false	true	9	Бородин Николай Александрович	1 979
							1 990
						вна	1 975
						вна	1 997
						вна	1 997
						вна	1 984
						вна	1 984
						вич	1 979
16	1	1	false	true	24	Руднев Максим Андреевич	1 990
17	1	1	false	true	33	Антипова Вероника Николавна	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.





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	1 975
2	1	1	false	true	35	Антипова Вероника Николаевна	1 975
3	2	1	false	true	9	Бородин Николай Александрович	1 979
4	2	1	false	true	19	Броодин Николай Александрович	1 979
5	1		false	true	6	Гасимова Юлия Ивагновна	1 985
6	1		false	true	2	Гасимова Юлия Ивановна	1 985
7	0			true	3	Запрудина Любовь Ивановна	1 992
8	0						1 992
9	1						1 997
10	1						1 997
11	0						1 975
12	0						1 975
13	1						1 990
14	1						1 990

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.