TÉCNICO LISBOA

Data Analysis and Integration

Approximate Duplicate Detection

Introduction

- String matching
  - Levenshtein, Damerau-Levenshtein, Needleman-Wunsch
  - Jaro, Jaro-Winkler
  - Jaccard
  - Soundex, Refined Soundex

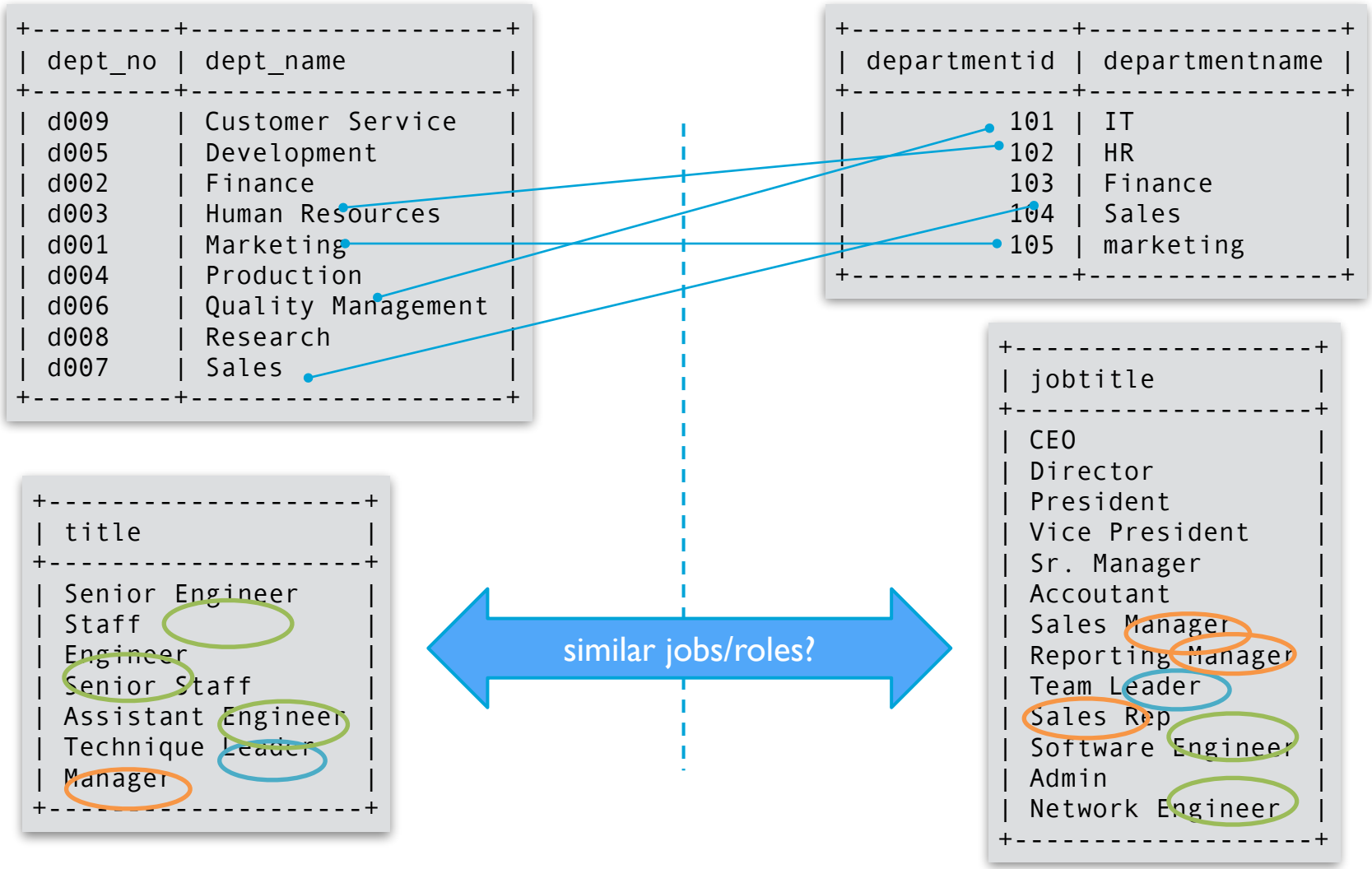
- Function that gives the similarity between two strings:

$$s(x, y) = \ldots$$

Data Matching

Data Matching

- How to find these matches?



Data Matching

- How to find these matches?

Find all combinations of strings that are similar

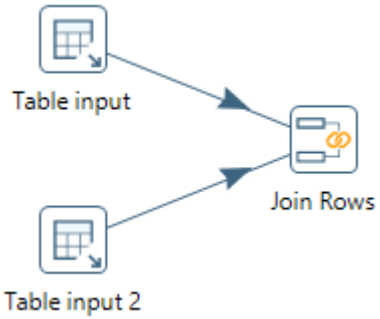
```
for each string x ∈ X do
    for each string y ∈ Y do
        if s(x, y) ≥ t then return (x, y) as a matched pair
    end for
end for
```

  - compare every x with every y
  - calculate similarity $s(x, y)$
  - use t as a threshold

Data Matching

Rows of step: Table input (9 rows)

#	dept_no	dept_name
1	d009	Customer Service
2	d005	Development
3	d002	Finance
4	d003	Human Resources
5	d001	Marketing
6	d004	Production
7	d006	Quality Management
8	d008	Research
9	d007	Sales



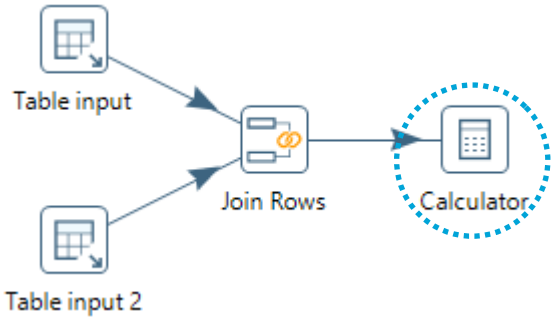
Rows of step: Table input 2 (5 rows)

#	departmentid	departmentname
1	101	IT
2	102	HR
3	103	Finance
4	104	Sales
5	105	marketing

Rows of step: Join Rows (45 rows)

#	dept_no	dept_name	departmentid	departmentname
1	d009	Customer Service	101	IT
2	d009	Customer Service	102	HR
3	d009	Customer Service	103	Finance
4	d009	Customer Service	104	Sales
5	d009	Customer Service	105	marketing
6	d005	Development	101	IT
7	d005	Development	102	HR
8	d005	Development	103	Finance
9	d005	Development	104	Sales
10	d005	Development	105	marketing
11	d002	Finance	101	IT
12	d002	Finance	102	HR
13	d002	Finance	103	Finance
14	d002	Finance	104	Sales
15	d002	Finance	105	marketing
16	d003	Human Resources	101	IT
17	d003	Human Resources	102	HR
18	d003	Human Resources	103	Finance
19	d003	Human Resources	104	Sales
20	d003	Human Resources	105	marketing
21	d001	Marketing	101	IT
22	d001	Marketing	102	HR
23	d001	Marketing	103	Finance
24	d001	Marketing	104	Sales
25	d001	Marketing	105	marketing
26	d004	Production	101	IT
27	d004	Production	102	HR

Data Matching



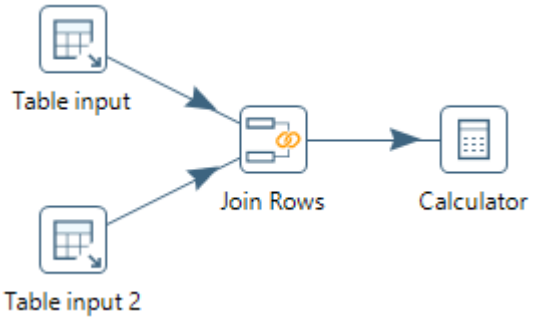
Rows of step: Calculator (45 rows)

#	dept_no	dept_name	departmentid	departmentname	levenshtein
1	d009	Customer Service	101	IT	16
2	d009	Customer Service	102	HR	16
3	d009	Customer Service	103	Finance	14
4	d009	Customer Service	104	Sales	15
5	d009	Customer Service	105	marketing	12
6	d005	Development	101	IT	11
7	d005	Development	102	HR	11
8	d005	Development	103	Finance	10
9	d005	Development	104	Sales	9
10	d005	Development	105	marketing	10
11	d002	Finance	101	IT	7
12	d002	Finance	102	HR	7
13	d002	Finance	103	Finance	0
14	d002	Finance	104	Sales	6
15	d002	Finance	105	marketing	9
16	d003	Human Resources	101	IT	15
17	d003	Human Resources	102	HR	13
18	d003	Human Resources	103	Finance	11
					12
					12
					9
					9
					9
24	d001	Marketing	104	Sales	7
25	d001	Marketing	105	marketing	1
26	d004	Production	101	IT	10
27	d004	Production	102	HR	10

Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type
1	levenshtein	Levenshtein Distance (source A and target B)	dept_name	departmentname		Integer

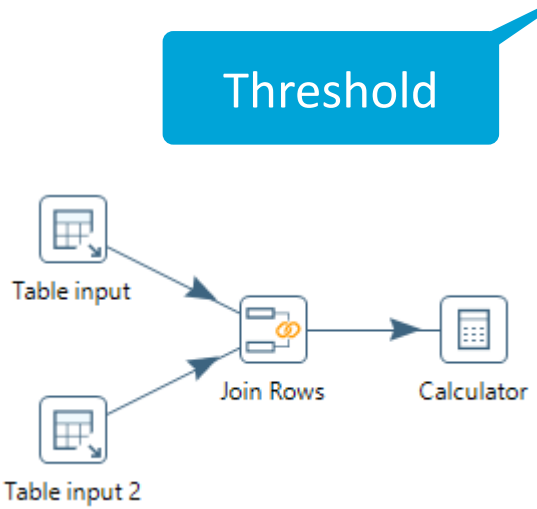
Data Matching



Rows of step: Calculator (45 rows)

#	dept_no	dept_name	departmentid	departmentname	levenshtein
1	d009	Customer Service	101	IT	16
2	d009	Customer Service	102	HR	16
3	d009	Customer Service	103	Finance	14
4	d009	Customer Service	104	Sales	15
5	d009	Customer Service	105	marketing	12
6	d005	Development	101	IT	11
7	d005	Development	102	HR	11
8	d005	Development	103	Finance	10
9	d005	Development	104	Sales	9
10	d005	Development	105	marketing	10
11	d002	Finance	101	IT	7
12	d002	Finance	102	HR	7
13	d002	Finance	103	Finance	0
14	d002	Finance	104	Sales	6
15	d002	Finance	105	marketing	9
16	d003	Human Resources	101	IT	15
17	d003	Human Resources	102	HR	13
18	d003	Human Resources	103	Finance	11
19	d003	Human Resources	104	Sales	12
20	d003	Human Resources	105	marketing	12
21	d001	Marketing	101	IT	9
22	d001	Marketing	102	HR	9
23	d001	Marketing	103	Finance	9
24	d001	Marketing	104	Sales	7
25	d001	Marketing	105	marketing	1
26	d004	Production	101	IT	10
27	d004	Production	102	HR	10

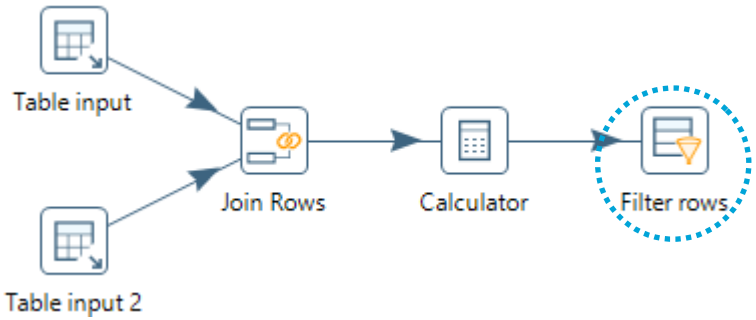
Data Matching

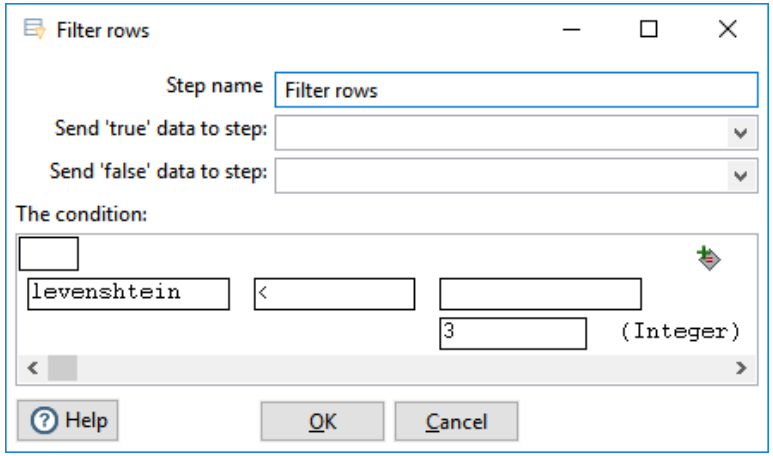


Rows of step: Calculator (45 rows)

#	dept_no	dept_name	departmentid	departmentname	levenshtein
13	d002	Finance	103	Finance	0
44	d007	Sales	104	Sales	0
25	d001	Marketing	105	marketing	1
41	d007	Sales	101	IT	5
42	d007	Sales	102	HR	5
14	d002	Finance	104	Sales	6
38	d008	Research	103	Finance	6
43	d007	Sales	103	Finance	6
11	d002	Finance	101	IT	7
12	d002	Finance	102	HR	7
24	d001	Marketing	104	Sales	7
39	d008	Research	104	Sales	7
45	d007	Sales	105	marketing	7
30	d004	Production	105	marketing	8
36	d008	Research	101	IT	8
37	d008	Research	102	HR	8
40	d008	Research	105	marketing	8
9	d005	Development	104	Sales	9
15	d002	Finance	105	marketing	9
21	d001	Marketing	101	IT	9
22	d001	Marketing	102	HR	9
23	d001	Marketing	103	Finance	9
28	d004	Production	103	Finance	9
8	d005	Development	103	Finance	10
10	d005	Development	105	marketing	10
26	d004	Production	101	IT	10
27	d004	Production	102	HR	10

Data Matching





Rows of step: Filter rows (3 rows)

#	dept_no	dept_name	departmentid	departmentname	levenshtein
1	d002	Finance	103	Finance	0
2	d001	Marketing	105	marketing	1
3	d007	Sales	104	Sales	0

Data Matching

- Potential problems
  - number of comparisons may be too large
    - e.g. 1000 x 1000 rows = 10^6 comparisons
  - difficult to find single perfect measure
    - e.g. there might be false negatives
  - there may be no clear-cut threshold
    - e.g. no way to avoid **false positives** or **false negatives**

Duplicate Detection

Duplicate Detection

- Similar approach for Duplicate Detection
  - multiple records refer to the same real-world entity

customer_id	first_name	last_name	address	city	country
169	Isabel	de Castro	Estrada da saúde n. 58	Lisboa	Portugal
504	Isabel	Castro	Estrada da saúde 58	Lisbon	Portugal
568	Isabella	de Castro	Estrada da saúde 58	Lisboa	Portugal
369	Manuel	Rodriguez	Jardim das Rosas 32	Lisboa	Portugal
535	Manuel	Rodrigues	Jardim das Rosas n. 32	Lisboa	Portugal
576	Emanuel	Rodrigues	Jardim das Rosas 32	Lisbon	Portugal

  - compare these records to find duplicates
    - focus on first_name, last_name, address

Duplicate Detection

- To find exact duplicates

```
select *
from customers as a, customers as b
where a.first_name = b.first_name
    and a.last_name = b.last_name
    and a.address = b.address;
```

But we want to find **approximate** duplicates

- where first_name, last_name, address are **similar**, not **equal** → **similar** according to some measure

Duplicate Detection

- We could write:

```
select *
from customers as a, customers as b
where similarity(a.first_name, b.first_name) >= …
    and similarity(a.last_name, b.last_name) >= …
    and similarity(a.address, b.address) >= …;
```

  - assuming that the ***similarity(…, …)*** function is available

Duplicate Detection

- Or:

```
select *
from customers as a, customers as b
where similarity(a.first_name, b.first_name) +
        similarity(a.last_name, b.last_name) +
        similarity(a.address, b.address) >= …;
```

  - using the similarities together in the same formula

Duplicate Detection

- Or even:

```
select *
from customers as a, customers as b
where 0.3*similarity(a.first_name, b.first_name) +
        0.3*similarity(a.last_name, b.last_name) +
        0.4*similarity(a.address, b.address) >= …;
```

  - using different weights for the similarities

General Approach

- Compute the **similarity** between two records x and y

$$s(x, y) = \sum_{i=1}^{n} \alpha_i \bullet s_i(x, y)$$

  - where:
    - $s_i(x, y)$ is the similarity between the i^{th} attributes of x and y
    - α_i is the weight given to the similarity $s_i(x, y)$

General Approach

- The goal is to find every match (x, y) for which

$$s(x, y) = \sum_{i=1}^{n} \alpha_i \bullet s_i(x, y) \geq t$$

  - where t is a threshold

Duplicate Detection

- Example

```
customer_id,first_name,last_name,address,city,country
169,Isabel,de Castro,Estrada da saúde n. 58,Lisboa,Portugal
504,Isabel,Castro,Estrada da saúde 58,Lisbon,Portugal
568,Isabella,de Castro,Estrada da saúde 58,Lisboa,Portugal
369,Manuel,Rodriguez,Jardim das Rosas 32,Lisboa,Portugal
535,Manuel,Rodrigues,Jardim das Rosas n. 32,Lisboa,Portugal
576,Emanuel,Rodrigues,Jardim das Rosas 32,Lisbon,Portugal
```

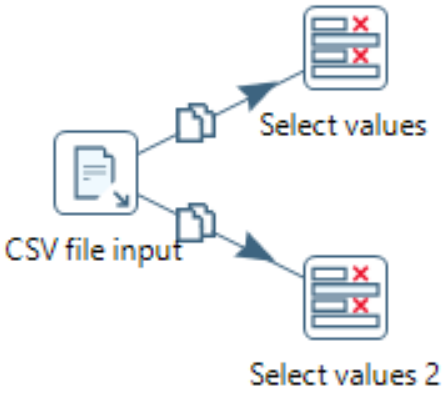
CSV file input

Rows of step: CSV file input (6 rows)

#	customer_id	first_name	last_name	address	city	country
1	169	Isabel	de Castro	Estrada da saúde n. 58	Lisboa	Portugal
2	504	Isabel	Castro	Estrada da saúde 58	Lisbon	Portugal
3	568	Isabella	de Castro	Estrada da saúde 58	Lisboa	Portugal
4	369	Manuel	Rodriguez	Jardim das Rosas 32	Lisboa	Portugal
5	535	Manuel	Rodrigues	Jardim das Rosas n. 32	Lisboa	Portugal
6	576	Emanuel	Rodrigues	Jardim das Rosas 32	Lisbon	Portugal

Duplicate Detection

- Rename



Rows of step: Select values (6 rows)

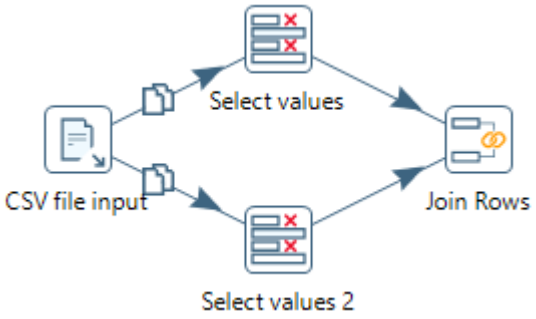
#	customer_id_1	first_name_1	last_name_1	address_1
1	169	Isabel	de Castro	Estrada da saúde n. 58
2	504	Isabel	Castro	Estrada da saúde 58
3	568	Isabella	de Castro	Estrada da saúde 58
4	369	Manuel	Rodriguez	Jardim das Rosas 32
5	535	Manuel	Rodrigues	Jardim das Rosas n. 32
6	576	Emanuel	Rodrigues	Jardim das Rosas 32

Rows of step: Select values 2 (6 rows)

#	customer_id_2	first_name_2	last_name_2	address_2
1	169	Isabel	de Castro	Estrada da saúde n. 58
2	504	Isabel	Castro	Estrada da saúde 58
3	568	Isabella	de Castro	Estrada da saúde 58
4	369	Manuel	Rodriguez	Jardim das Rosas 32
5	535	Manuel	Rodrigues	Jardim das Rosas n. 32
6	576	Emanuel	Rodrigues	Jardim das Rosas 32

Duplicate Detection

- Compare



Rows of step: Join Rows (36 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2
1	169	Isabel	de Castro	Estrada da saúde n. 58	169	Isabel	de Castro	Estrada da saúde n. 58
2	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58
3	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58
4	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32
5	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
6	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
7	504	Isabel	Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
8	504	Isabel	Castro	Estrada da saúde 58	504	Isabel	Castro	Estrada da saúde 58
9	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58
10	504	Isabel	Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32
11	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
12	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
13	568	Isabella	de Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
14	568	Isabella	de Castro	Estrada da saúde 58	504	Isabel	Castro	Estrada da saúde 58
15	568	Isabella	de Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58
16	568	Isabella	de Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32

Duplicate Detection

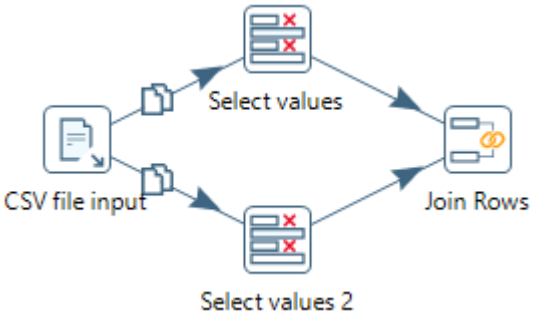
- No need to compare when $x \leftrightarrow x$

Rows of step: Join Rows (36 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2
1	169	Isabel	de Castro	Estrada da saúde n. 58	169	Isabel	de Castro	Estrada da saúde n. 58
2	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58
3	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58
4	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32
5	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
6	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
7	504	Isabel	Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
8	504	Isabel	Castro	Estrada da saúde 58	504	Isabel	Castro	Estrada da saúde 58
9	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58
10	504	Isabel	Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32
11	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
12	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
13	568	Isabella	de Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
14	568	Isabella	de Castro	Estrada da saúde 58	504	Isabel	Castro	Estrada da saúde 58
15	568	Isabella	de Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58
16	568	Isabella	de Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32

Duplicate Detection

- No need to compare when $x \leftrightarrow x$



The condition:

customer_id_1 <> customer_id_2

Rows of step: Join Rows (30 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2
1	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58
2	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58
3	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32
4	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
5	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
6	504	Isabel	Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
7	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58
8	504	Isabel	Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32
9	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
10	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
11	568	Isabella	de Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
12	568	Isabella	de Castro	Estrada da saúde 58	504	Isabel	Castro	Estrada da saúde 58
13	568	Isabella	de Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32
14	568	Isabella	de Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
15	568	Isabella	de Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
16	369	Manuel	Rodriguez	Jardim das Rosas 32	169	Isabel	de Castro	Estrada da saúde n. 58

Duplicate Detection

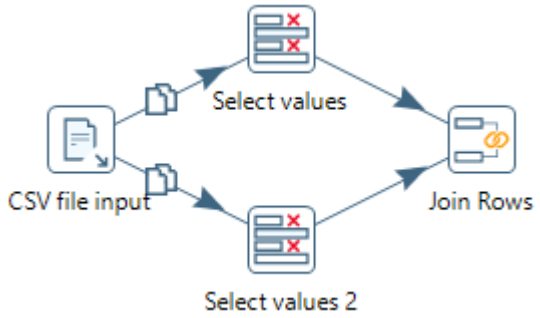
- No need to compare both $x \leftrightarrow y$ and $y \leftrightarrow x$

Rows of step: Join Rows (30 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2
1	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58
2	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58
3	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32
4	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
5	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
6	504	Isabel	Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
7	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58
8	504	Isabel	Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32
9	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
10	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
11	568	Isabella	de Castro	Estrada da saúde 58	169	Isabel	de Castro	Estrada da saúde n. 58
12	568	Isabella	de Castro	Estrada da saúde 58	504	Isabel	Castro	Estrada da saúde 58
13	568	Isabella	de Castro	Estrada da saúde 58	369	Manuel	Rodriguez	Jardim das Rosas 32
14	568	Isabella	de Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
15	568	Isabella	de Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
16	369	Manuel	Rodriguez	Jardim das Rosas 32	169	Isabel	de Castro	Estrada da saúde n. 58

Duplicate Detection

- No need to compare both $x \leftrightarrow y$ and $y \leftrightarrow x$



The condition:

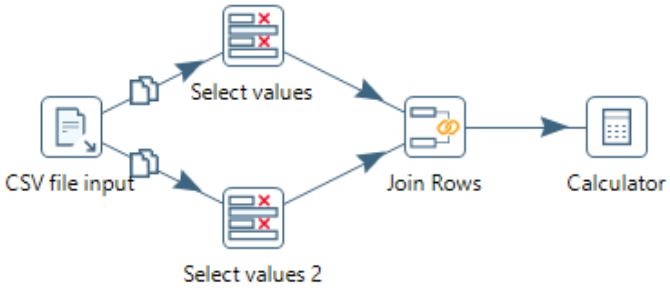
customer_id_1 < customer_id_2

Rows of step: Join Rows (15 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2
1	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58
2	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58
3	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32
4	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
5	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
6	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58
7	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32
8	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
9	568	Isabella	de Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32
10	369	Manuel	Rodriguez	Jardim das Rosas 32	504	Isabel	Castro	Estrada da saúde 58
11	369	Manuel	Rodriguez	Jardim das Rosas 32	568	Isabella	de Castro	Estrada da saúde 58
12	369	Manuel	Rodriguez	Jardim das Rosas 32	535	Manuel	Rodrigues	Jardim das Rosas n. 32
13	369	Manuel	Rodriguez	Jardim das Rosas 32	576	Emanuel	Rodrigues	Jardim das Rosas 32
14	535	Manuel	Rodrigues	Jardim das Rosas n. 32	568	Isabella	de Castro	Estrada da saúde 58
15	535	Manuel	Rodrigues	Jardim das Rosas n. 32	576	Emanuel	Rodrigues	Jardim das Rosas 32

Duplicate Detection

- Calculating the measures



Fields:

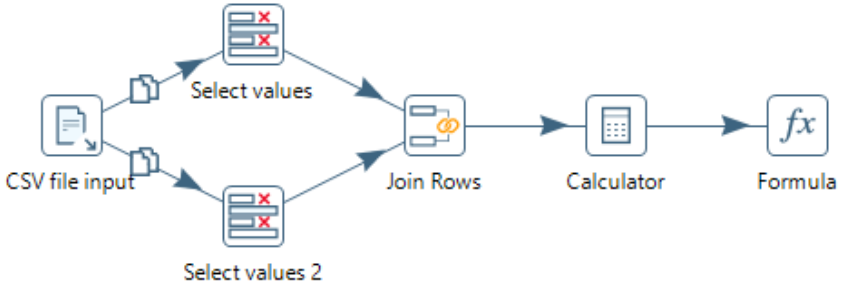
#	New field	Calculation	Field A	Field B	Value type
1	sim1	Levenshtein Distance (source A and target B)	first_name_1	first_name_2	Number
2	sim2	Levenshtein Distance (source A and target B)	last_name_1	last_name_2	Number
3	sim3	Levenshtein Distance (source A and target B)	address_1	address_2	Number

Rows of step: Calculator (15 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2	sim1	sim2	sim3
1	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58	0.0	3.0	3.0
2	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	0.0	3.0
3	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32	4.0	9.0	16.0
4	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32	4.0	9.0	14.0
5	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	4.0	9.0	16.0
6	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	3.0	0.0
7	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32	4.0	9.0	15.0
8	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	4.0	9.0	14.0
9	568	Isabella	de Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	6.0	9.0	14.0
10	369	Manuel	Rodriguez	Jardim das Rosas 32	504	Isabel	Castro	Estrada da saúde 58	4.0	9.0	14.0
11	369	Manuel	Rodriguez	Jardim das Rosas 32	568	Isabella	de Castro	Estrada da saúde 58	6.0	9.0	14.0
12	369	Manuel	Rodriguez	Jardim das Rosas 32	535	Manuel	Rodrigues	Jardim das Rosas n. 32	0.0	1.0	3.0
13	369	Manuel	Rodriguez	Jardim das Rosas 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	1.0	0.0
14	535	Manuel	Rodrigues	Jardim das Rosas n. 32	568	Isabella	de Castro	Estrada da saúde 58	6.0	9.0	15.0
15	535	Manuel	Rodrigues	Jardim das Rosas n. 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	0.0	3.0

Duplicate Detection

- Combining the measures



Fields:

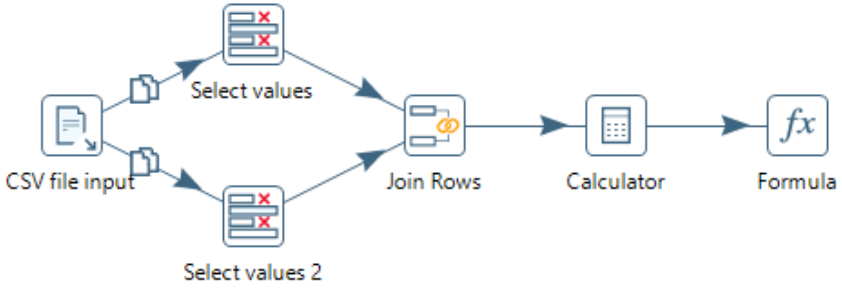
#	New field	Formula	Value type
1	sim_total	0.3*[sim1]+0.3*[sim2]+0.4*[sim3]	Number

Rows of step: Formula (15 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2	sim1	sim2	sim3	sim_total
1	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58	0.0	3.0	3.0	2.1
2	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	0.0	3.0	1.8
3	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32	4.0	9.0	16.0	10.3
4	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32	4.0	9.0	14.0	9.5
5	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	4.0	9.0	16.0	10.3
6	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	3.0	0.0	1.5
7	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32	4.0	9.0	15.0	9.9
8	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	4.0	9.0	14.0	9.5
9	568	Isabella	de Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	6.0	9.0	14.0	10.1
10	369	Manuel	Rodriguez	Jardim das Rosas 32	504	Isabel	Castro	Estrada da saúde 58	4.0	9.0	14.0	9.5
11	369	Manuel	Rodriguez	Jardim das Rosas 32	568	Isabella	de Castro	Estrada da saúde 58	6.0	9.0	14.0	10.1
12	369	Manuel	Rodriguez	Jardim das Rosas 32	535	Manuel	Rodrigues	Jardim das Rosas n. 32	0.0	1.0	3.0	1.5
13	369	Manuel	Rodriguez	Jardim das Rosas 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	1.0	0.0	0.9
14	535	Manuel	Rodrigues	Jardim das Rosas n. 32	568	Isabella	de Castro	Estrada da saúde 58	6.0	9.0	15.0	10.5
15	535	Manuel	Rodrigues	Jardim das Rosas n. 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	0.0	3.0	1.8

Duplicate Detection

- Combining the measures



Fields:

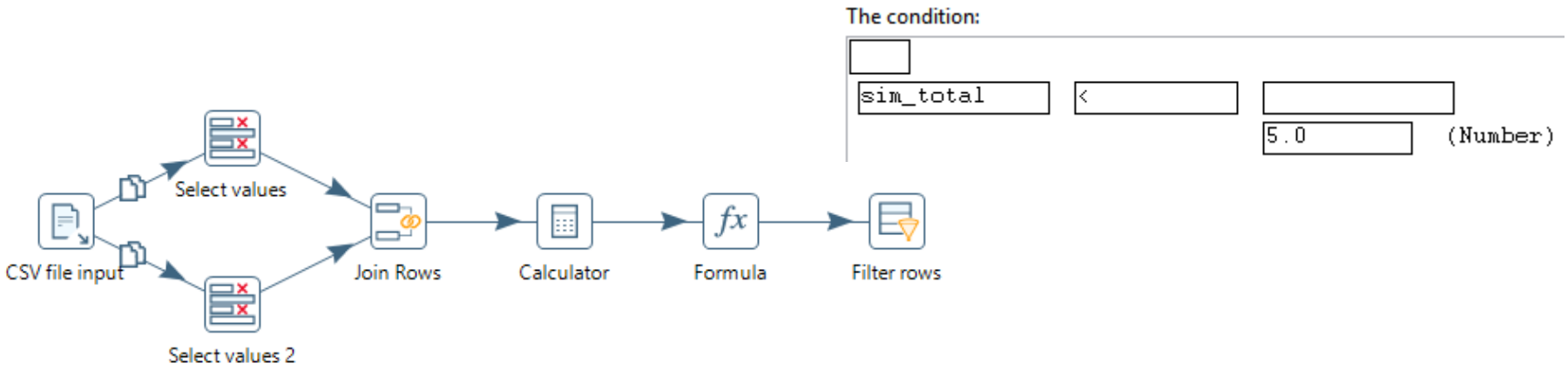
#	New field	Formula	Value type
1	sim_total	0.3*[sim1]+0.3*[sim2]+0.4*[sim3]	Number

Rows of step: Formula (15 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2	sim1	sim2	sim3	sim_total
13	369	Manuel	Rodriguez	Jardim das Rosas 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	1.0	0.0	0.9
6	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	3.0	0.0	1.5
12	369	Manuel	Rodriguez	Jardim das Rosas 32	535	Manuel	Rodrigues	Jardim das Rosas n. 32	0.0	1.0	3.0	1.5
2	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	0.0	3.0	1.8
15	535	Manuel	Rodrigues	Jardim das Rosas n. 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	0.0	3.0	1.8
1	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58	0.0	3.0	3.0	2.1
4	169	Isabel	de Castro	Estrada da saúde n. 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32	4.0	9.0	14.0	9.5
8	504	Isabel	Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	4.0	9.0	14.0	9.5
10	369	Manuel	Rodriguez	Jardim das Rosas 32	504	Isabel	Castro	Estrada da saúde 58	4.0	9.0	14.0	9.5
7	504	Isabel	Castro	Estrada da saúde 58	535	Manuel	Rodrigues	Jardim das Rosas n. 32	4.0	9.0	15.0	9.9
9	568	Isabella	de Castro	Estrada da saúde 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	6.0	9.0	14.0	10.1
11	369	Manuel	Rodriguez	Jardim das Rosas 32	568	Isabella	de Castro	Estrada da saúde 58	6.0	9.0	14.0	10.1
3	169	Isabel	de Castro	Estrada da saúde n. 58	369	Manuel	Rodriguez	Jardim das Rosas 32	4.0	9.0	16.0	10.3
5	169	Isabel	de Castro	Estrada da saúde n. 58	576	Emanuel	Rodrigues	Jardim das Rosas 32	4.0	9.0	16.0	10.3
14	535	Manuel	Rodrigues	Jardim das Rosas n. 32	568	Isabella	de Castro	Estrada da saúde 58	6.0	9.0	15.0	10.5

Duplicate Detection

- Applying the threshold

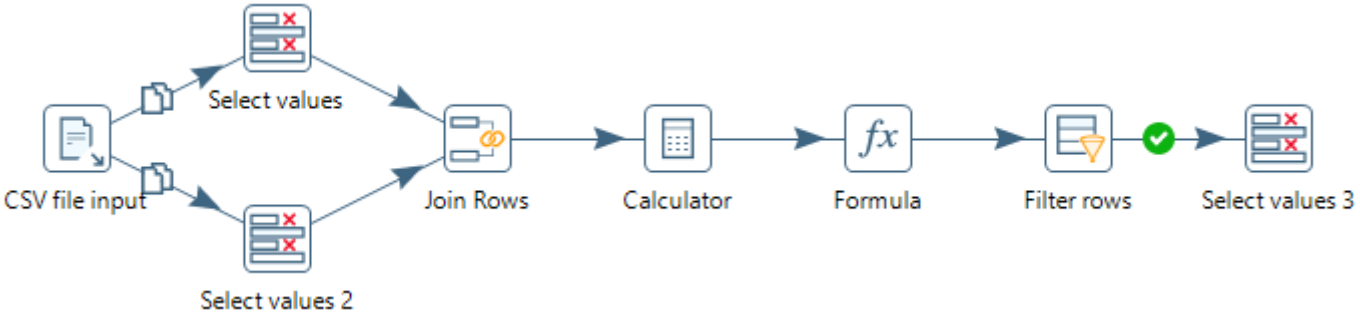


Rows of step: Filter rows (6 rows)

#	customer_id_1	first_name_1	last_name_1	address_1	customer_id_2	first_name_2	last_name_2	address_2	sim1	sim2	sim3	sim_total
1	169	Isabel	de Castro	Estrada da saúde n. 58	504	Isabel	Castro	Estrada da saúde 58	0.0	3.0	3.0	2.1
2	169	Isabel	de Castro	Estrada da saúde n. 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	0.0	3.0	1.8
3	504	Isabel	Castro	Estrada da saúde 58	568	Isabella	de Castro	Estrada da saúde 58	2.0	3.0	0.0	1.5
4	369	Manuel	Rodriguez	Jardim das Rosas 32	535	Manuel	Rodrigues	Jardim das Rosas n. 32	0.0	1.0	3.0	1.5
5	369	Manuel	Rodriguez	Jardim das Rosas 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	1.0	0.0	0.9
6	535	Manuel	Rodrigues	Jardim das Rosas n. 32	576	Emanuel	Rodrigues	Jardim das Rosas 32	2.0	0.0	3.0	1.8

Duplicate Detection

- Getting the results



Rows of step: Select values 3 (6 rows)

#	customer_id_1	customer_id_2
1	169	504
2	169	568
3	504	568
4	369	535
5	369	576
6	535	576

Duplicate Detection

- Using the results

Rows of step: CSV file input (6 rows)

#	customer_id	first_name	last_name	address	city	country
1	169	Isabel	de Castro	Estrada da saúde n. 58	Lisboa	Portugal
2	504	Isabel	Castro	Estrada da saúde 58	Lisbon	Portugal
3	568	Isabella	de Castro	Estrada da saúde 58	Lisboa	Portugal
4	369	Manuel	Rodriguez	Jardim das Rosas 32	Lisboa	Portugal
5	535	Manuel	Rodrigues	Jardim das Rosas n. 32	Lisboa	Portugal
6	576	Emanuel	Rodrigues	Jardim das Rosas 32	Lisbon	Portugal

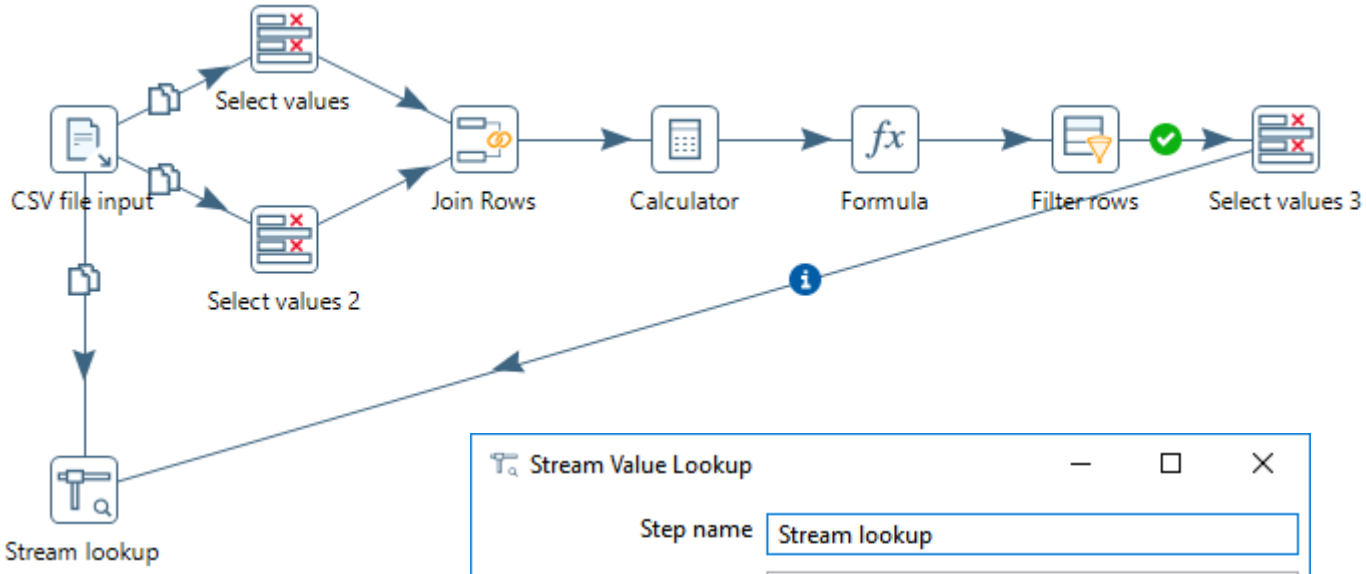
Once we have the table with the matches we can go back and…

Rows of step: Select values 3 (6 rows)

#	customer_id_1	customer_id_2
1	169	504
2	169	568
3	504	568
4	369	535
5	369	576
6	535	576

Duplicate Elimination

Duplicate Elimination



… look them up

Stream Value Lookup

Step name: Stream lookup

Lookup step: Select values 3

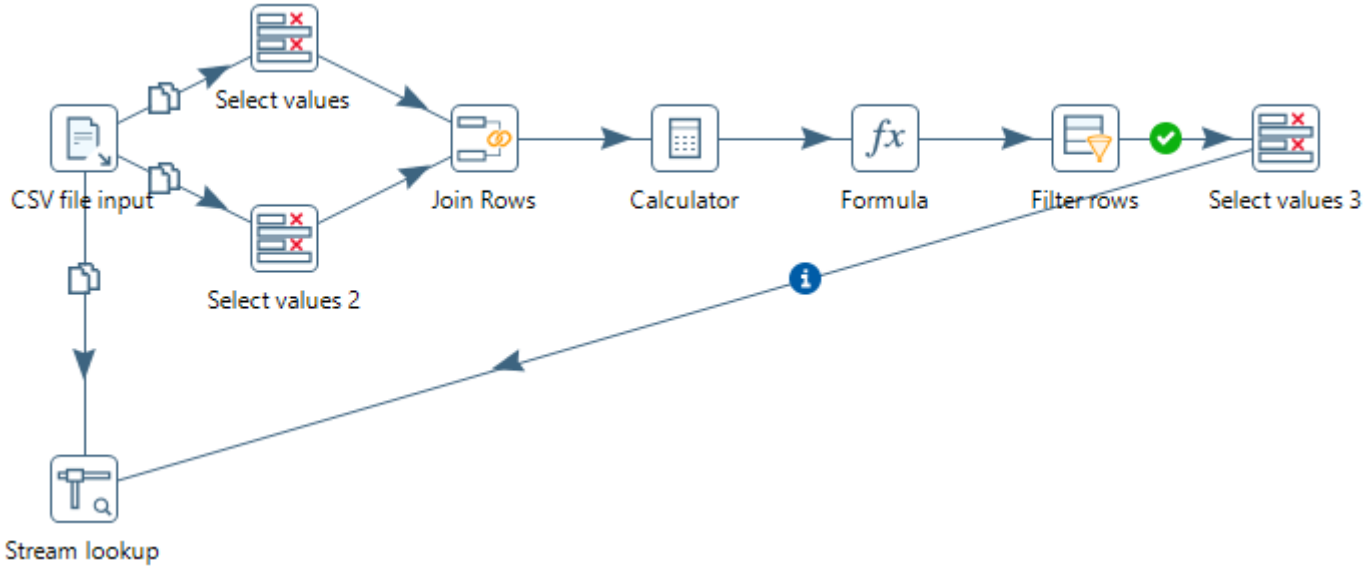
The key(s) to look up the value(s):

#	Field	LookupField
1	customer_id	customer_id_2

Specify the fields to retrieve :

#	Field	New name	Default	Type
1	customer_id_1	duplicate		Integer

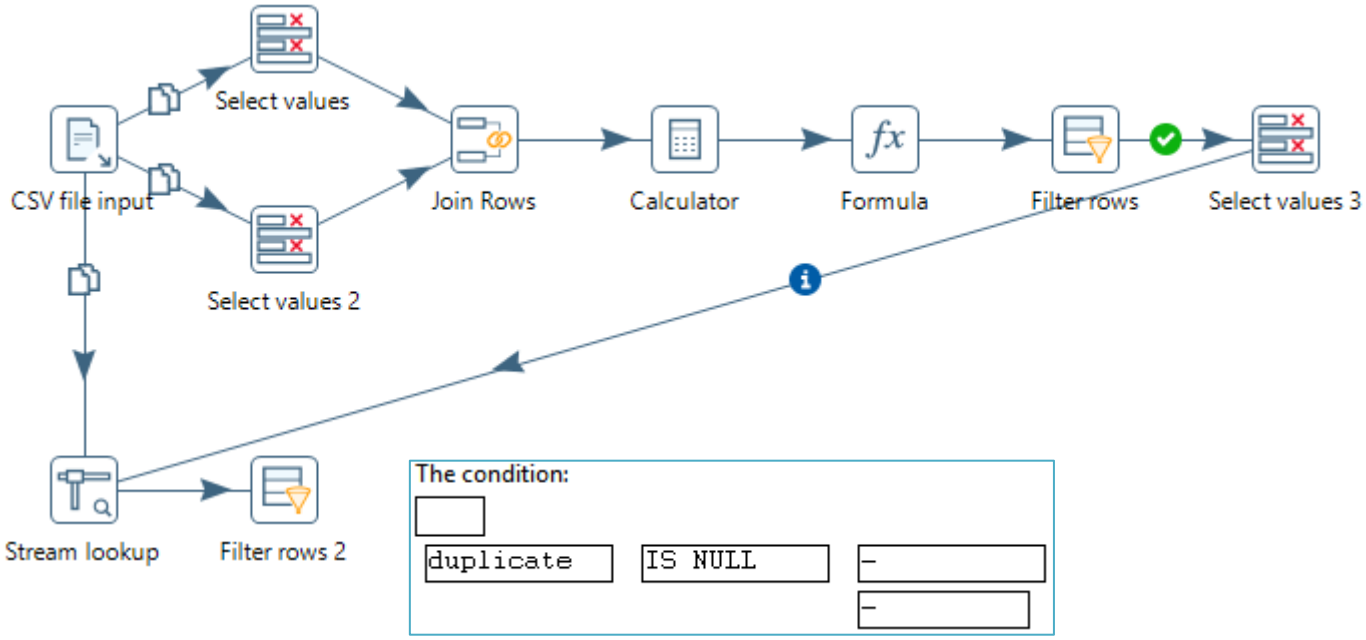
Duplicate Elimination



Rows of step: Stream lookup (6 rows)

#	customer_id	first_name	last_name	address	city	country	duplicate
1	169	Isabel	de Castro	Estrada da saúde n. 58	Lisboa	Portugal	<null>
2	504	Isabel	Castro	Estrada da saúde 58	Lisbon	Portugal	169
3	568	Isabella	de Castro	Estrada da saúde 58	Lisboa	Portugal	504
4	369	Manuel	Rodriguez	Jardim das Rosas 32	Lisboa	Portugal	<null>
5	535	Manuel	Rodrigues	Jardim das Rosas n. 32	Lisboa	Portugal	369
6	576	Emanuel	Rodrigues	Jardim das Rosas 32	Lisbon	Portugal	535

Duplicate Elimination



Rows of step: Filter rows 2 (2 rows)

#	customer_id	first_name	last_name	address	city	country	duplicate
1	169	Isabel	de Castro	Estrada da saúde n. 58	Lisboa	Portugal	<null>
2	369	Manuel	Rodriguez	Jardim das Rosas 32	Lisboa	Portugal	<null>