

Real Time Fuzzy Matching With Spark and ElasticSearch

About Us

The only way to do great work is to love what you do.

- Steve Jobs



The problem - lake or swamp?





Duplicates



Challenges

- Quadratic problem
- No standard notion of similarity
- Omissions, typos and other issues
- Different languages



Use Case - Customer Record Dedup



Use Case - Customer Record Dedup



Use Case - Shopping Site Comparison





Use Case - Shopping Site Comparison





Other Use Cases

- Cross selling
- Financial Credit Ratings
- Fraud Analytics
- Catalog and inventory management
- Household and individual level analytics.



Lets start wishing...

- Data variety
- Scalable
- No manual configuration of rules or algorithms
- Multi language
- Real time



Jobs Support Logout Do the following pairs match? Unsure 0/0 Non Match 0/0 Match 0/0 brooklyn blake o'connell street eden hills 810 vic 19101201 2258884 brooklyn jehne dumaresq street rowethorpe lamington 6061 wa 19420723 2935590 Can't Say caitlin noack hamilton row sec 375 manly vale 7011 vic 8328282 caitlin westbrook 631 blacksdane private sanctuary gidgegannup 3114 19360908 9439473 casson street nsw Can't Say



Jobs Support Logout

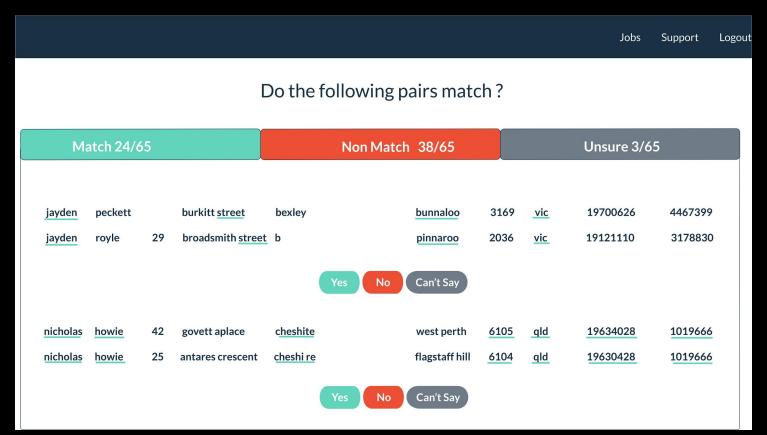
Do the following pairs match?





Jobs Support Logout Do the following pairs match? Match 10/35 Unsure 1/35 Non Match 24/35 3500 6806487 isabella keira street pyndari toowoomba vic 19400405 isabella hingston 57 loureiro street thornlie 3500 19290125 4688367 nsw Can't Say wilkey 1399646 nicholas 12 caladenia street rivonia kingston 2480 wa wuikikakiwi 2126121 nicholas wickel 12 jardine street highton 3085 19570620 wa Can't Say







Real Time

Spark + ElasticSearch

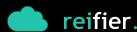




Spark Benefits

- Distributed
- Scalable
- Fast
- Machine Learning
- Sampling
- No need to orchestrate multiple jobs





Thank You!

www.nubetech.co

sonal@nubetech.co

