



# You Belong Together

## Detecting Linked Accounts at Ricardo

By Tobias Kaymak





Data Engineering



Data Science



# Ricardo

Ricardo

+ Sell Item  
Free

Favourites

tobias.kaymak

EN

Search for products, sellers or product number



All categories



Antiques & artwork



Vehicles



Manual work & gardening



Household & home furnishings



Clothing & accessories



Hobbies & model building



Sports



Watches & jewellery

## Hol dir den #RicardoBus!

Egal ob für ein gemütliches Familien-Wochenende oder die Party mit Freunden: Mach dich bereit für unvergessliche Sommertage.

JETZT ENTDECKEN



## Currently on Ricardo

Nearby

Ending soon

From CHF 1.-

Popular

New

## Our picks for you



BMC Timemachine Road Disc

28 Jul. 2021, 19:32

Starting bid 3'150.00

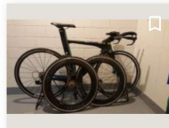
Buy now 3'500.00



BMC time machine TM01

1 Aug. 2021, 06:13

Buy now 2'200.00



BMC Timemachine TM01

25 Jul. 2021, 12:26

Buy now 3'500.00



Triathlonvelo Merida Time Warp Größe M

24 Jul. 2021, 20:02

Starting bid 2'500.00

Buy now 3'500.00



BMC Timemachine TM02 Ultegra Di2

28 Jul. 2021, 16:53

Buy now 3'100.00



BEAM  
SUMMIT

Are the Olson Twins...



...the same person?



# Data Sources



Historical data  
in BigQuery

# Switzerland is a multi-language country





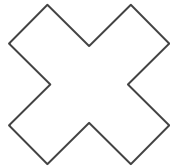
# Fuzzy String Matching

Quellgasse

Quellgase

Quelgasse

...



Quellgasse

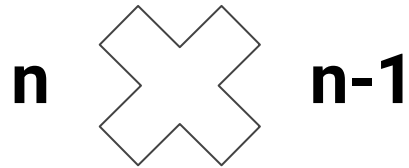
Quellgase

Quelgasse

...



# Fuzzy String Matching



# Phonetic Algorithms

Quellgase

Quellgasse    Quelgasse



# Cologne phonetics



# Cologne phonetics

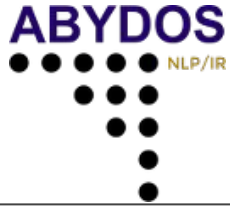
**458**

Quellgase

Quellgasse    Quelgasse

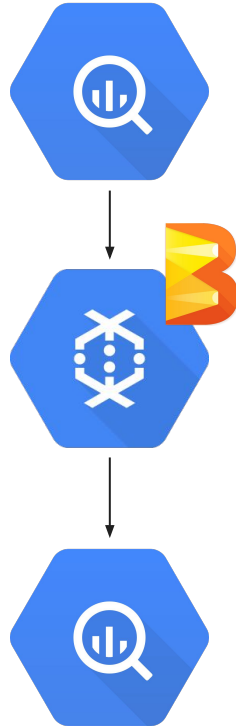


# Can we solve it with Python?



458

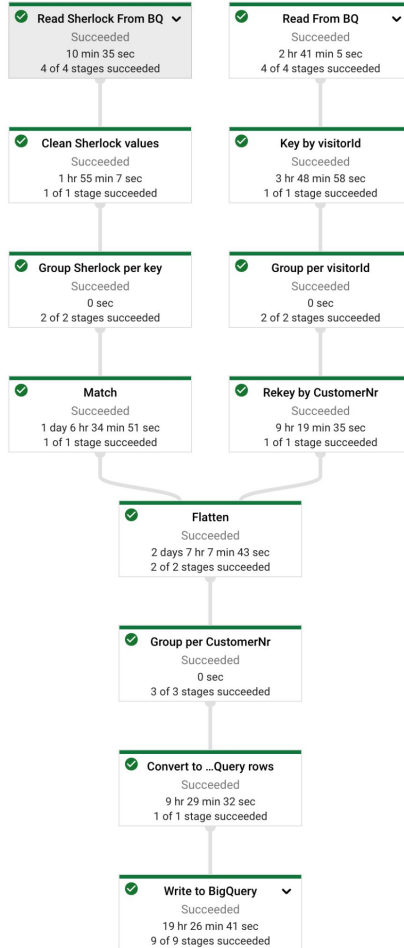
Quellgase  
Quellgasse Quellgasse



```
with beam.Pipeline() as p:
```

```
p  
| "Read data" >> beam.io.ReadFromBigQuery()  
  
| "Encode phonetic" >> beam.FlatMap(enc_values)  
| "Group per phonetic key" >> beam.GroupByKey()  
| "Match values" >> beam.FlatMap(match_values)  
| "Group per customerNr" >> beam.GroupByKey()  
| "Convert to BigQuery rows" >> beam.Map(to_row)  
  
| "Write result" >> beam.io.WriteToBigQuery()
```





So are they the same person?





# The Beam Python Experience is Fun



Apache  
Airflow



# The Price of Context Switching?

Data Engineering



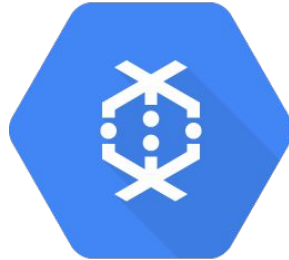
Data Science



# Lessons learned



# Lessons learned



`--experiments=use_runner_v2`

`--flexrs_goal=COST_OPTIMIZED`



Thank you ❤️



# References

- <https://pypi.org/project/abydos/>
- <https://beam.apache.org/documentation/dsls/dataframes/overview/>
- <https://issues.apache.org/jira/browse/BEAM-11998>
- <https://issues.apache.org/jira/browse/BEAM-11993>
- <https://issues.apache.org/jira/browse/BEAM-11991>
- <https://cloud.google.com/composer/docs/how-to/using/using-dataflow-template-operator>

Images:

John Oliver / Last Week Tonight / HBO

<https://www.shutterstock.com/image-photo/mary-kate-olsen-ashley-womens-wear-185453450>

<https://www.newlyswissed.com/official-languages-of-switzerland/>

(CC-BY-SA 4.0) - Roland Zumbuehl

<https://upload.wikimedia.org/wikipedia/commons/b/b4/2005-Biel-Quellgasse.jpg>

