You Belong Together

Detecting Linked Accounts at Ricardo

By Tobias Kaymak

BEAM SUMMIT

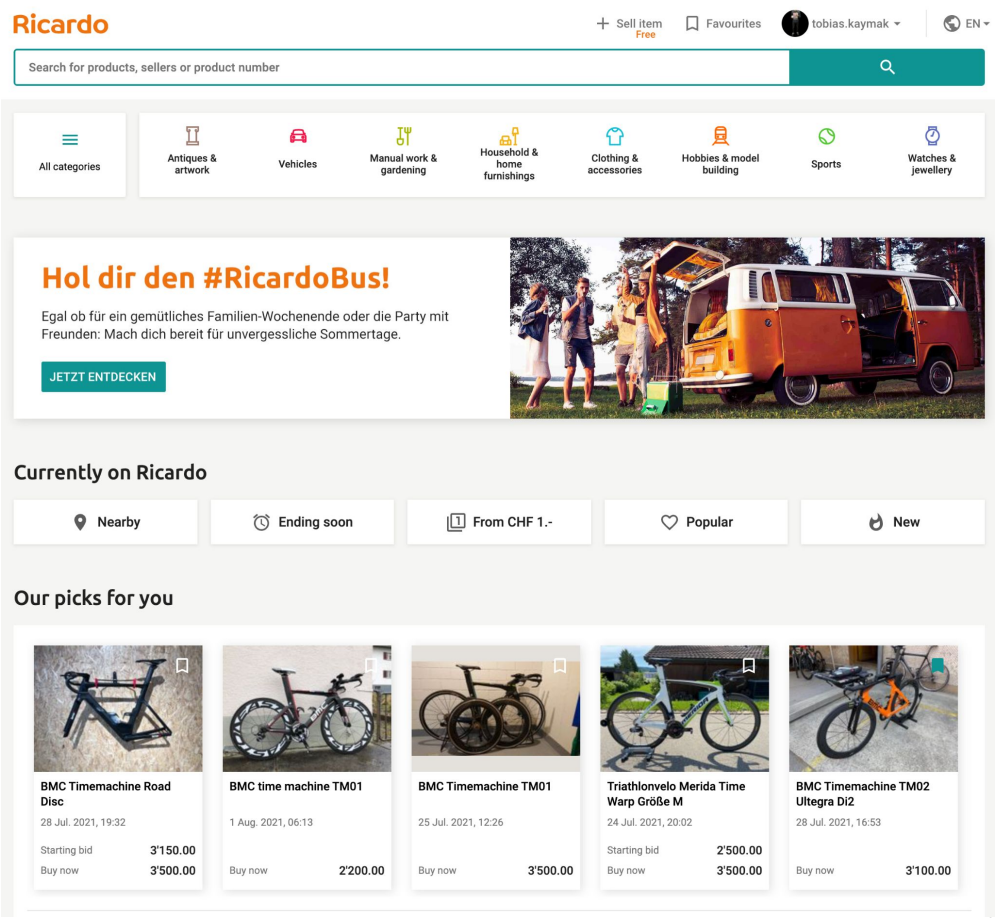
BEAM SUMMIT

Data Engineering

Data Science

BEAM SUMMIT

Ricardo



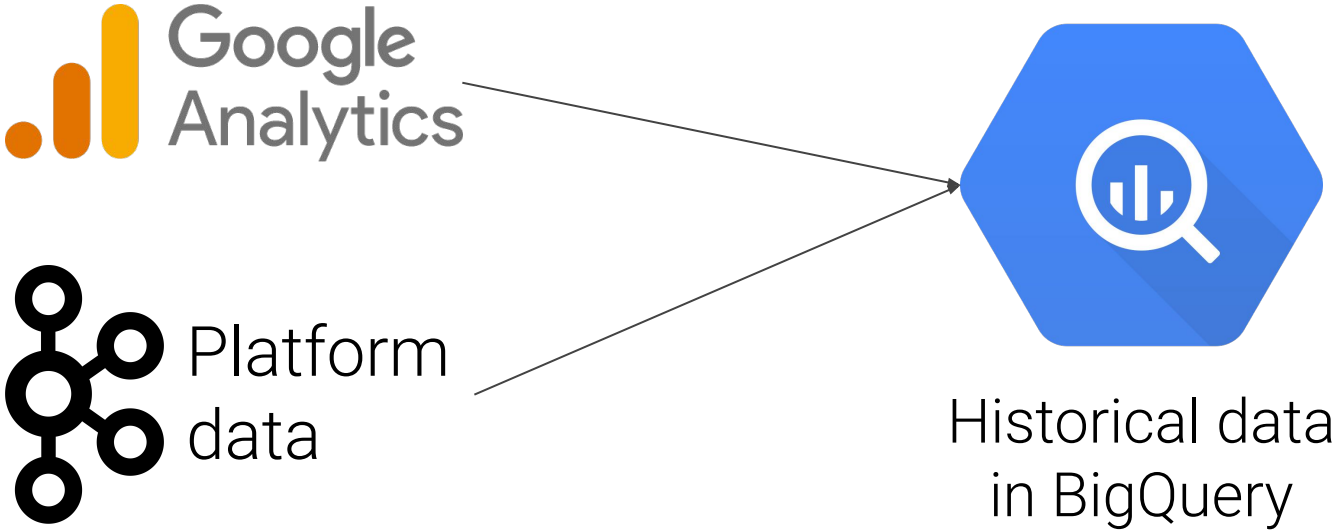
BEAM SUMMIT

Are the Olson Twins...

BEAM SUMMIT

...the same person?

Data Sources



Switzerland is a multi-language country

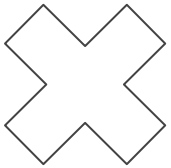
Fuzzy String Matching

Quellgasse

Quellgase

Quelgasse

...



Quellgasse

Quellgase

Quelgasse

...

Fuzzy String Matching

n ✕ **n-1**

Phonetic Algorithms

Quellgase

Quellgasse Quelgasse

Cologne phonetics

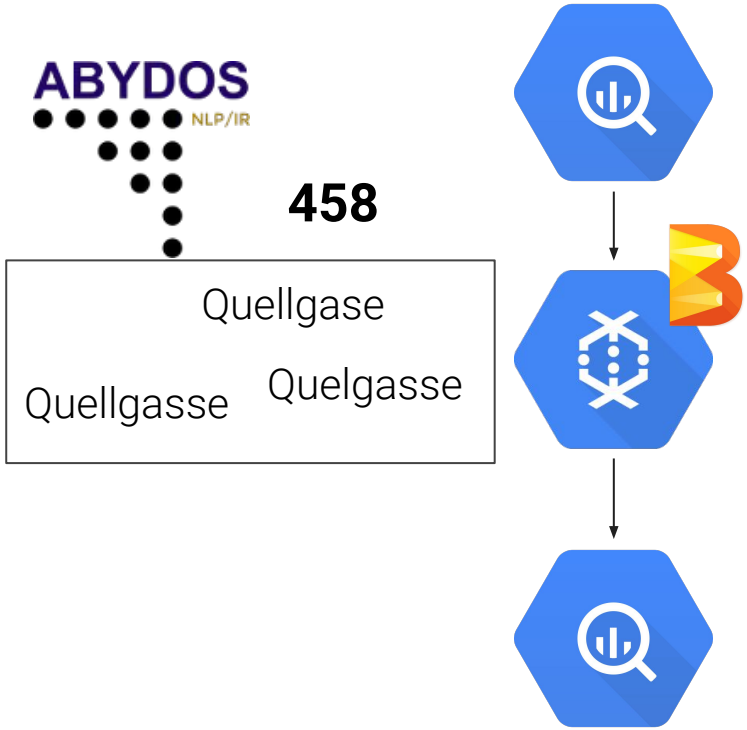
Cologne phonetics

458

Quellgase

Quellgasse Quelgasse

Can we solve it with Python?



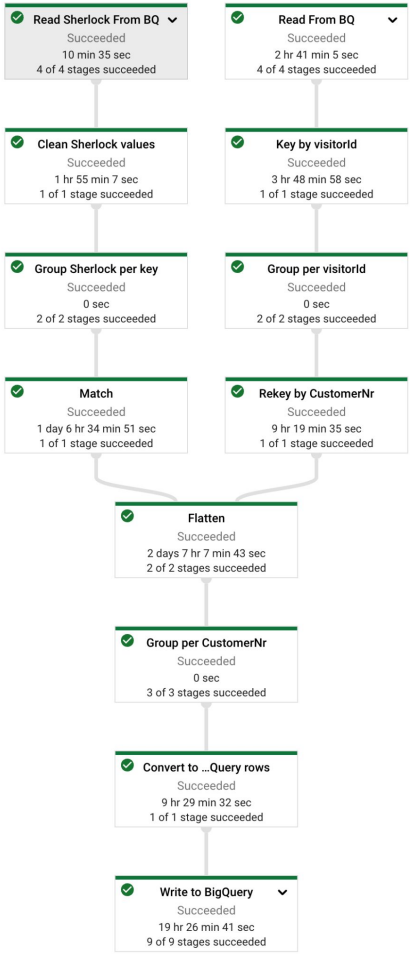

```
with beam.Pipeline() as p:
    p
    | "Read data" >> beam.io.ReadFromBigQuery()

    | "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()
```

BEAM SUMMIT


python
TM


Read Sherlock From BQ
Succeeded
10 min 35 sec
4 of 4 stages succeeded
Read From BQ
Succeeded
2 hr 41 min 5 sec
4 of 4 stages succeeded
Clean Sherlock values
Succeeded
1 hr 55 min 7 sec
1 of 1 stage succeeded
Key by visitorId
Succeeded
3 hr 48 min 58 sec
1 of 1 stage succeeded
Group Sherlock per key
Succeeded
0 sec
2 of 2 stages succeeded
Group per visitorId
Succeeded
0 sec
2 of 2 stages succeeded
Match
Succeeded
1 day 6 hr 34 min 51 sec
1 of 1 stage succeeded
Rekey by CustomerNr
Succeeded
9 hr 19 min 35 sec
1 of 1 stage succeeded
Flatten
Succeeded
2 days 7 hr 7 min 43 sec
2 of 2 stages succeeded
Group per CustomerNr
Succeeded
0 sec
3 of 3 stages succeeded
Convert to ...Query rows
Succeeded
9 hr 29 min 32 sec
1 of 1 stage succeeded
Write to BigQuery
Succeeded
19 hr 26 min 41 sec
9 of 9 stages succeeded


BEAM
SUMMIT

So are they the same person?

The Beam Python Experience is Fun

The Price of Context Switching?

Data Engineering

Data Science

Lessons learned

BEAM SUMMIT

Lessons learned

```
--experiments=use_runner_v2

--flexrs_goal=COST_OPTIMIZED
```

Thank you ❤️

BEAM SUMMIT

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

BEAM SUMMIT