Konzeption und Umsetzung einer DSL zur Informationsfusion auf verteilten heterogenen Graphen

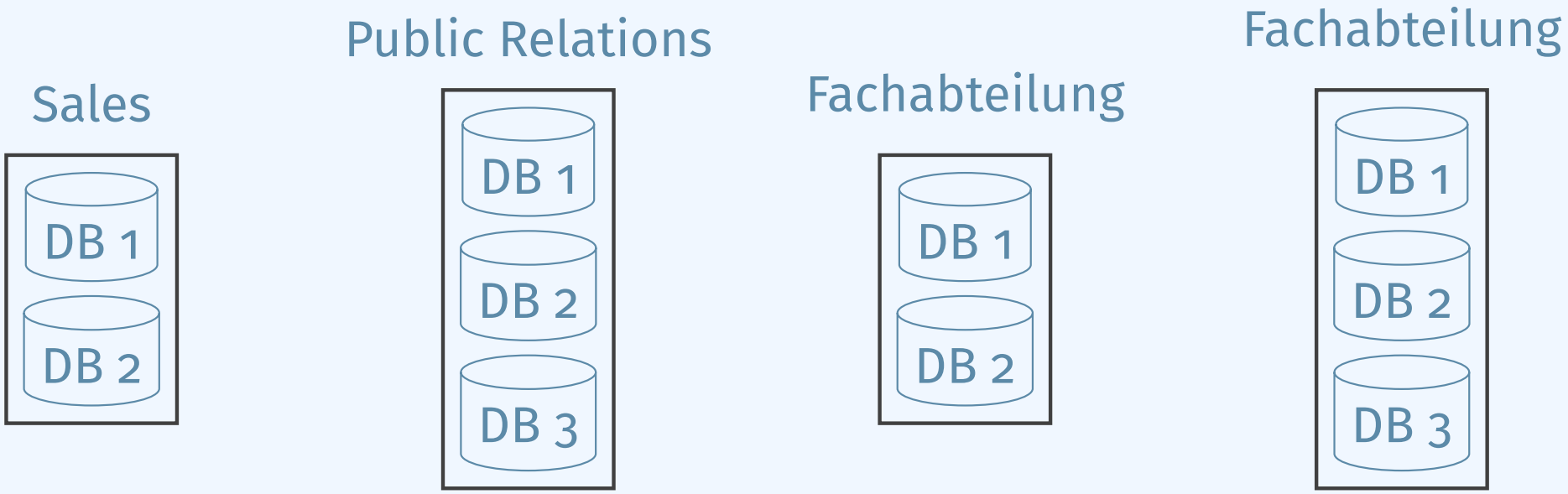
Alexander Kern

BTW 2019

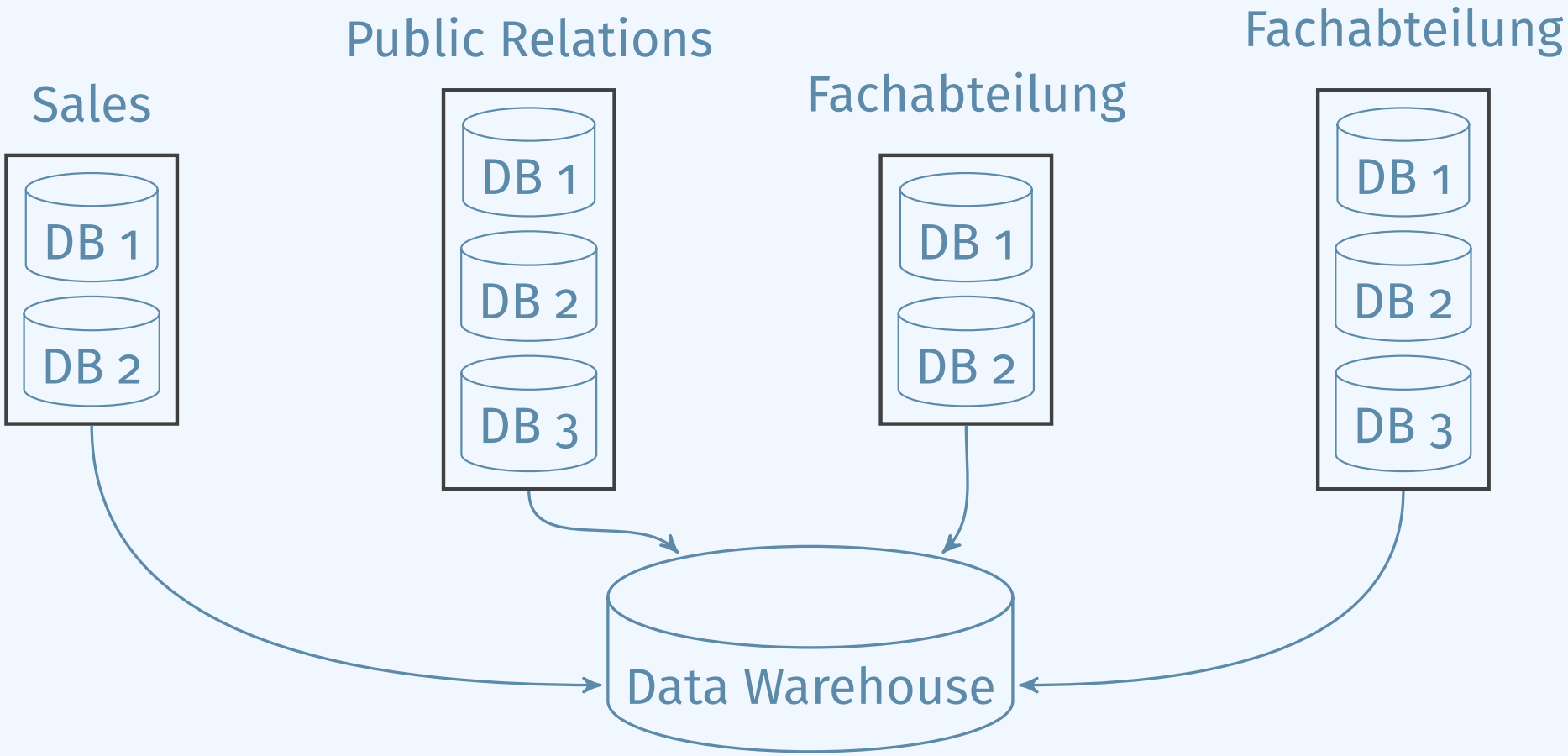
05.03.2019

Informationsfusion

Informationsfusion



Informationsfusion



Beispiele

- Verknüpfung verschiedener Datenbanken
- Aggregation von Suchergebnissen
- Sensormesswerte

Ziele

- Neue Zusammenhänge
- Wissen präzisieren
- Fehler finden
- Lücken füllen

Probleme

- Entity Resolution
- Duplikatserkennung
- Schemaintegration
- Informationsfusion

Probleme

- Entity Resolution
- Deduplikation
- Schemaintegration
- Informationsfusion

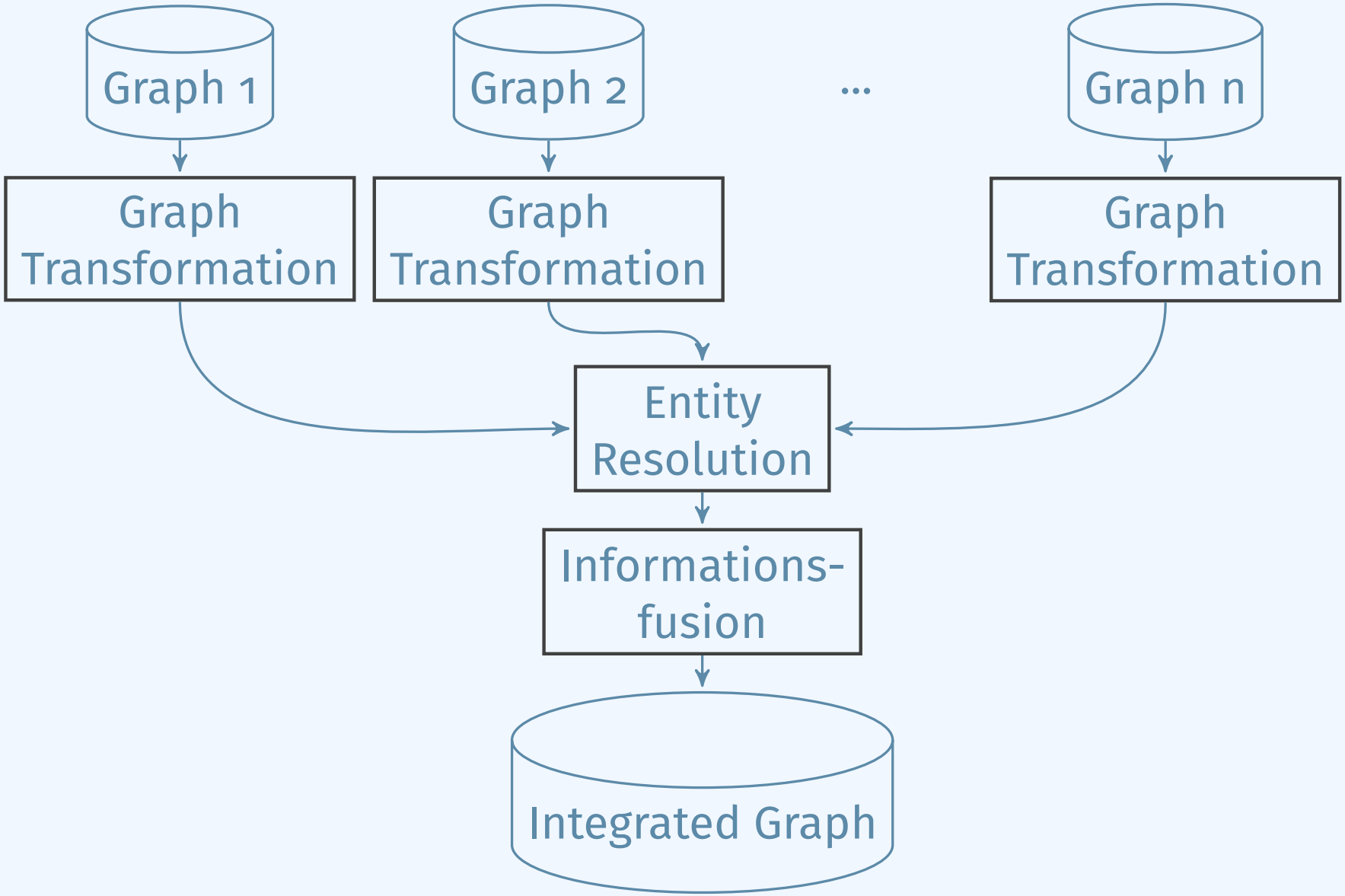
Probleme - Beispiele

Modellschema	Adresse	Straße, Nummer, PLZ, Ort
	name=Alexander	name=Kern
Inhaltliche Fehler	Augustusplatz 10	Ritterstrasse 9-13
	Augustusplatz 10	Augsutusplatz 10
Datenschema	2003-01-07	7. Jan. 2018

Gradoop

- Apache Flink-basiertes Framework
- Verteilte deklarative Graphanalysen
- Extended Property Graph Model

Prozess in Gradoop



Anforderungen

- Unterstützung heterogener Datenschema
- Beschreibung des Ausgabeschemas
- Zusammenhang mit Eingabedaten
- Bestimmung der Konfliktlösungsstrategien

  $\rightarrow$ Transformationsregeln

Domänenspezifische Sprache

Verwandte Arbeiten

- Declarative Data Cleaning: Language, Model, and Algorithms, Helena Galhardas et. al., 2001
- SQL-ähnliche Sprache zur Beschreibung von Informationsintegration
- Datenaufbereitung, Clustering und Merging

Verwandte Arbeiten

```
CREATE MERGING MergeAuthors
USING clusterAuthors(cluster id) ca
LET name = getLongestAuthorName(
        DirtyAuthors(ca).name)
key = generateKey()
{ SELECT key AS authorKey,
        name AS name INTO Authors g }
```

Design - Anforderungen

- Auswahl von (heterogenen) Eingabeattributen
- Strategien mit Optionen
- Kurzformen:
  - homogene Eingabeattribute
  - Attribute ohne Regeln

Design - Anforderungen

- Ausführung mit Gradoop/Flink
- Deklarativ

Attributwertkonfliktlösung

Strategie	Beschreibung
Straight	Erster Nicht-Null-Wert nach Ordnung
Retain	Konkatenation
Priority	Erster Wert in Ordnung der Quellen
Newest	Aktuellster Zeitstempel
Majority	Meistauftretender Wert
Source	Höchstes summiertes Gewicht
Property	Auswahl anhand der vorhandenen Attributwerte

Vergleiche [2]

Design - SQL

```
CREATE MERGING
LET Author
  name = property(
    (Source1.author.name,
     Source2.author.full_name),
        textual:longest),
  authorKey = straight(key)
```

Design - Kotlin

```
node Author {
  property name {
    Source1.author.name
    Source2.author.full_name
  } strategy property { textual:longest }

  property authorKey { key } strategy straight
}
```

Design - YML

```
nodes:
  - label: Person
    transformations:
      - outputProperty: name
        strategy:
          type: property
          options: [textual-longest]
        properties:
          - Source1.author.name
          - Source2.author.full_name
      - outputProperty: authorKey
        strategy:
          type: straight
        properties: [key]
```

Design - Evaluation

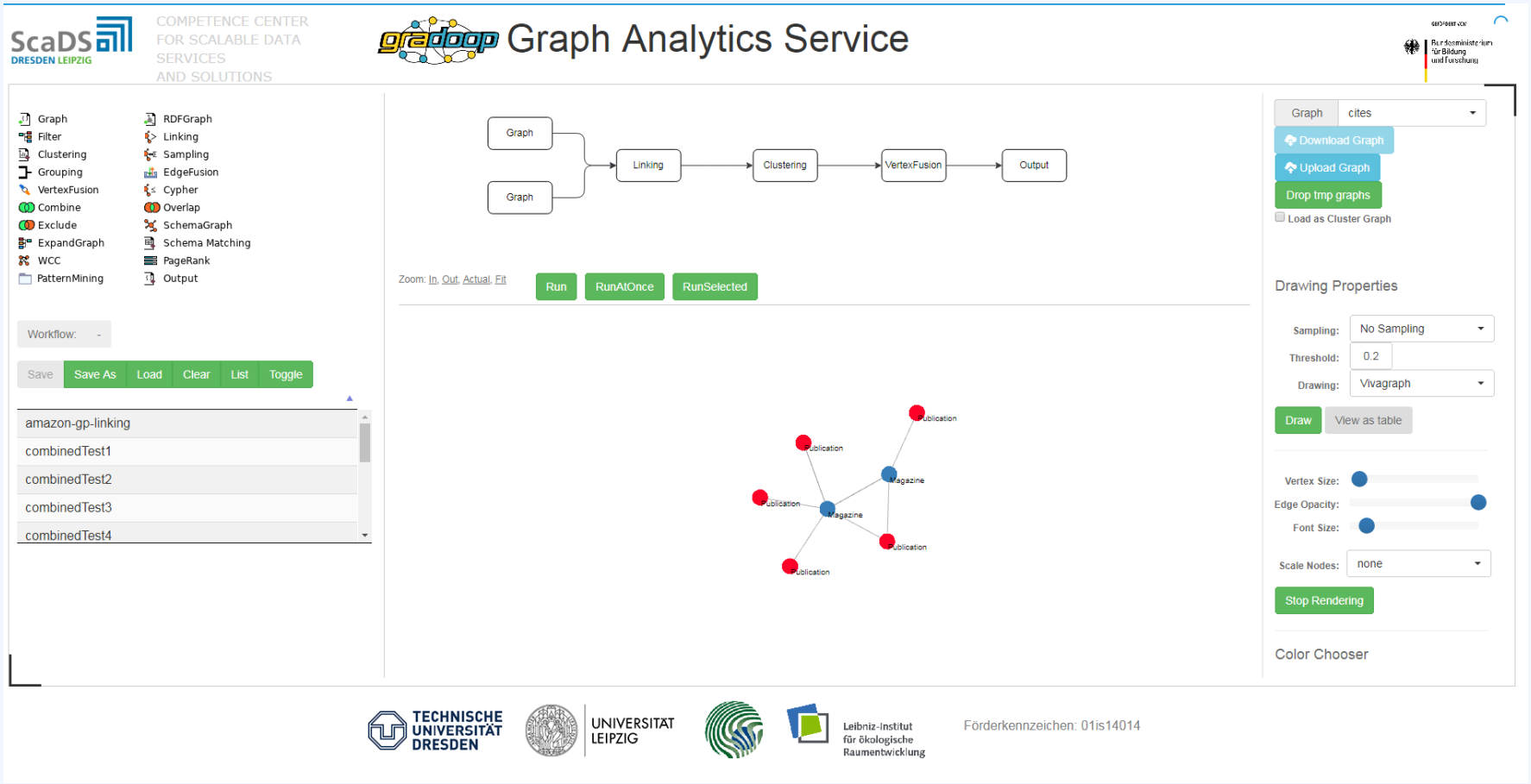
- Vorkenntnisse, Bekanntheit
- Toolunterstützung
- Einarbeitungszeit, Komplexität
- Abbildung der Problemdomäne
- ...

Design - Evaluation

- Vorkenntnisse
- Bekanntheit
- Einarbeitungszeit

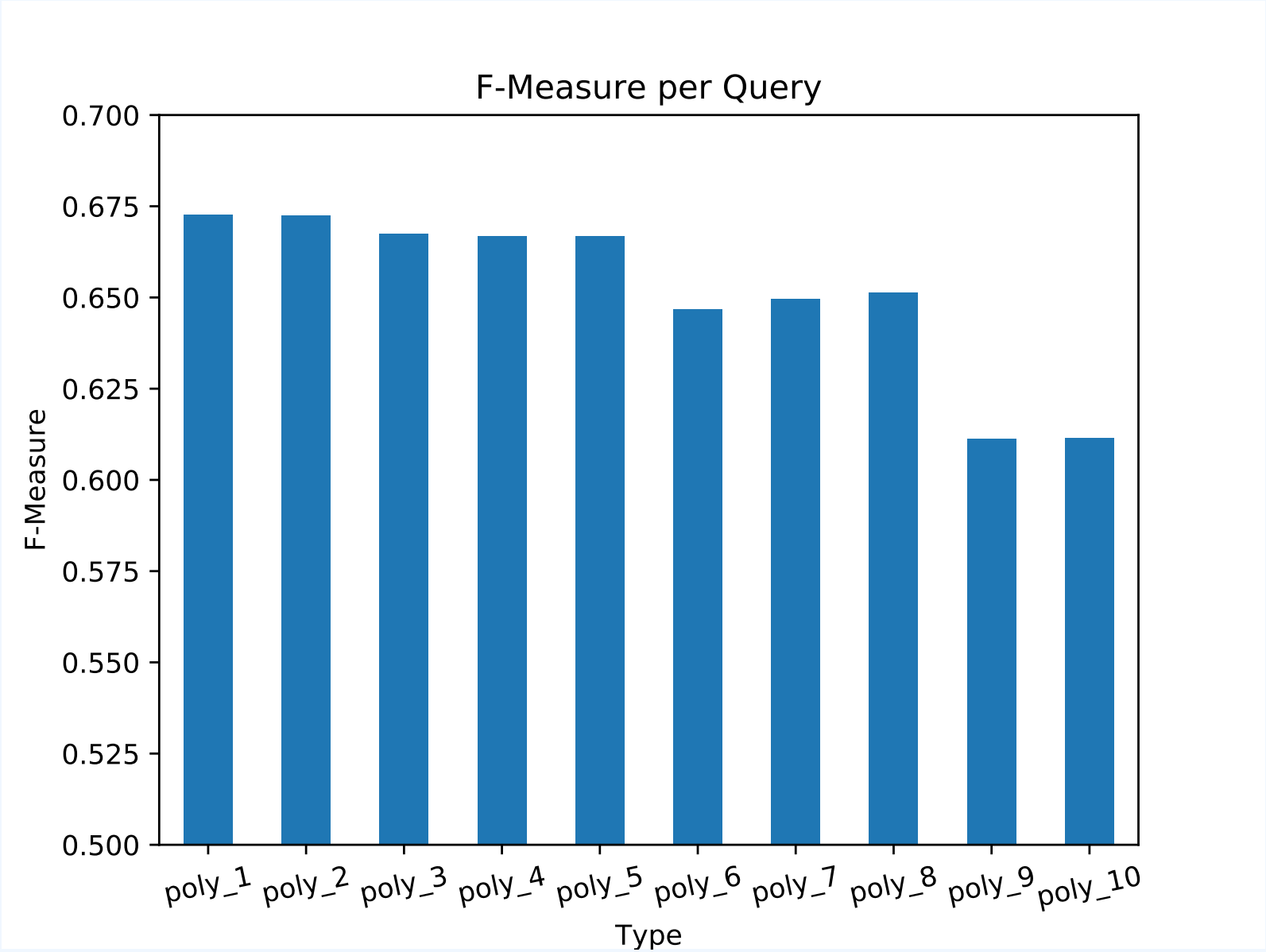
  $\rightarrow$ SQL-basierte DSL

Ergebnis - Gradoop Service

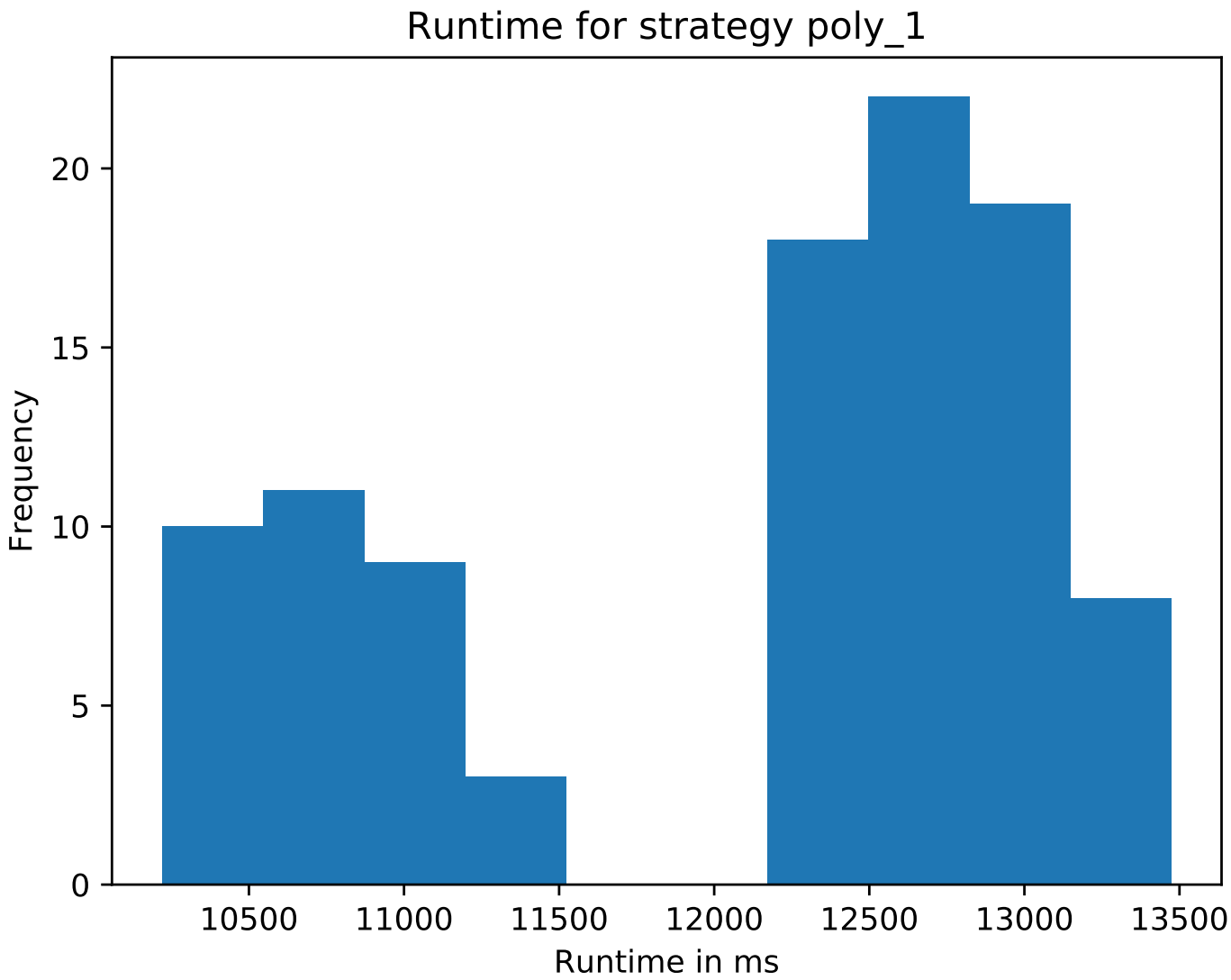


Evaluation

Evaluation - Matching Qualität



Evaluation - Laufzeit



Evaluation - Skalierung

Parallelisierung	64	32	16	8	4	2
Mittelwert	35.29	35.02	47.12	75.49	155.7	271.06
Median	35.46	34.42	47.30	75.31	156.62	272.57
Faktor	-	0.97	1.37	1.59	2.08	1.74

Fazit und Ausblick

Fazit

- Beispielhafte DSL-Implementierung
- In Code und Service nutzbar

Ausblick

- Evaluation mit Endnutzern
- Erweiterung für Kanten
- Weitere Konfliktlösungsstrategien
- Zusätzliche Anwendungsfälle

Vielen Dank für ihre Aufmerksamkeit!

References

Helena Galhardas, Daniela Florescu, and Dennis Shasha.
Declarative data cleaning: Language, model, and algorithms.
In *In VLDB*, pages 371–380, 2001.

Felix Naumann and Matthias Häussler.
Declarative data merging with conflict resolution.
In *International Conference on Information Quality (IQ 2002). 2002*, pages 212–224, 2002.