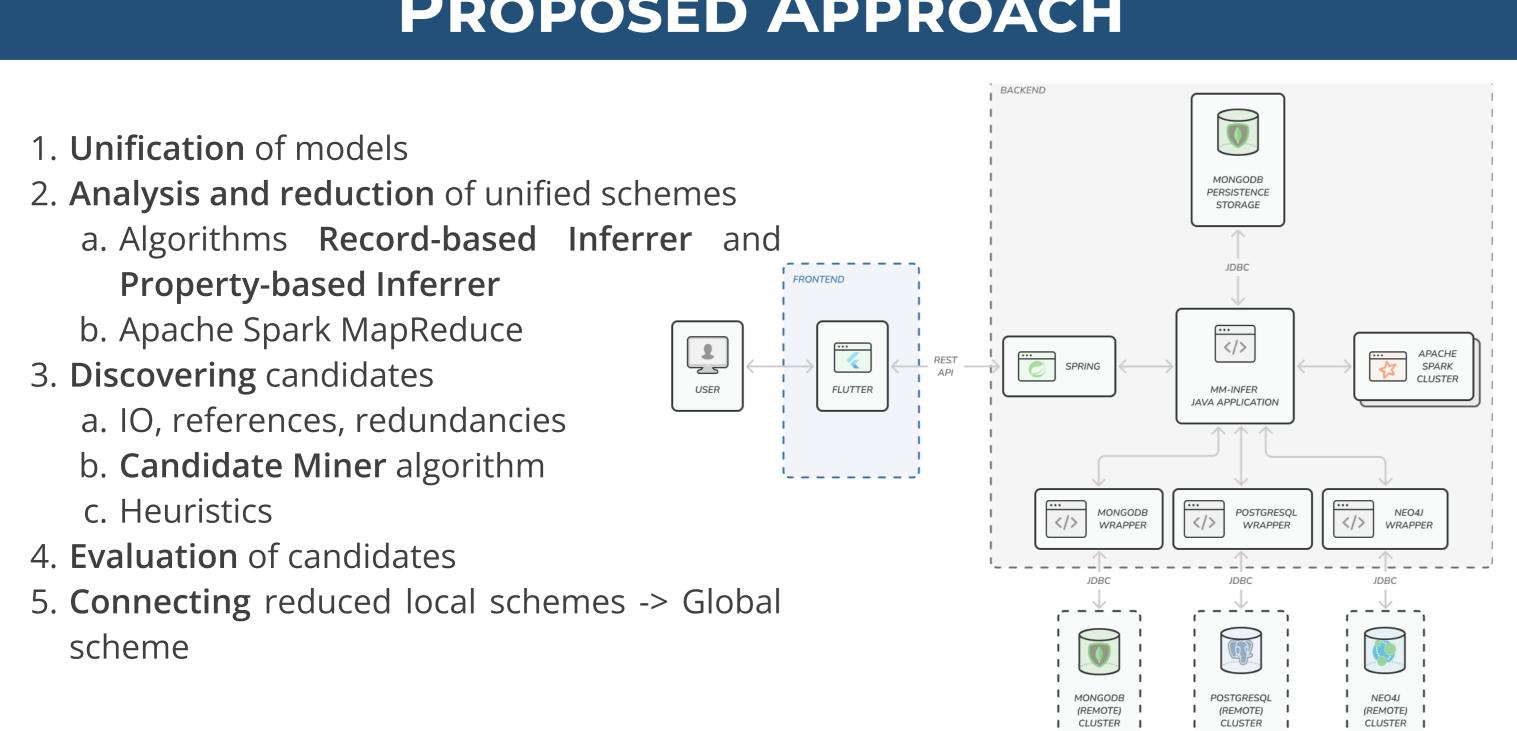
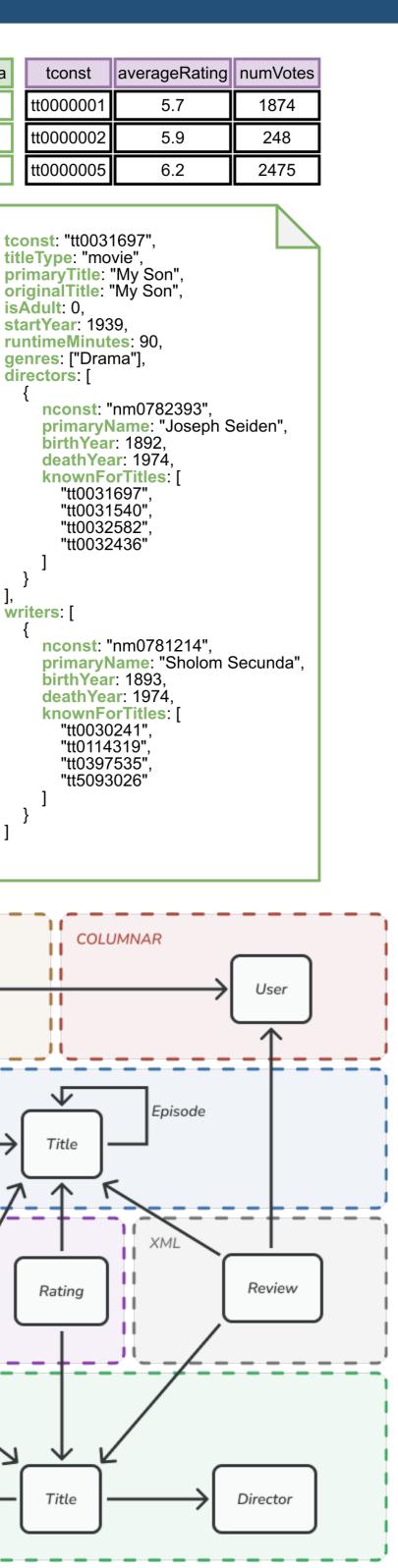
SCHEMA INFERENCE FOR MULTI-MODEL DATA

Author: Mgr. Sebastián Hricko, Supervisor: doc. RNDr. Irena Holubová, Ph.D., Advisor: Ing. Pavel Koupil Department of Software Engineering, Faculty of Mathematics and Physics, Charles University

PROBLEM DEFINI	τιον
primaryName: "Joseph Seiden", birthYear: 1892, deathYear: 1974,titleIdorderingtitlebirthYear: 1974, primaryProfession: ["producer", "director", "cinematographer"], knownForTitles: ["tt0031697", "tt0031540", "tt0032582", "tt0032436"]titleIdorderingtitlett0000054Blacksmith Sceler tt006812011The New Price Is I	ne
isAdult: 0, startYear: 1972,	ritePeople 0000001", 0000058"]
Episode titleType: "tvSeries", primaryTitle: "The Price Is Right", originalTitle: "The New Price Is Right", isAdult: 0, startYear: 2000, runtimeMinutes: 60, genres: ["Family", "Game-Show", "Reality-TV"] Title userId username userId email uu0091591 John W.	:00
UUU000003 Rose 1985	ouriteGenres 'Comedy" } vouritePeople
 Consider a multi-model scenario, where the data storage model corresponds to the expected use of the data (Document, Key-Value, Columnar, Relational, Graph,) Questions we aim to answer: What is the schema of the data in each 	n Principal
• We want to obtain a global view of the data	Additional

PROPOSED APPROACH





RECORD-BASED INFERRER

- 1. Traversing collections by **records** (i.e. PostgreSQL table rows)
- created
- collection

tconst	averageRating	numVotes
tt0000001	5.7	1874
tt0000002	5.9	1874

Rating RSDs
(_, U, (1, 1), U, [Map], [REL], [(tconst, U, (1, 1), U, [String], [F (averageRating, U, (1, 1), U, [C (numVotes, U, (1, 1), U, [Long]], ε , ε)
(U (1 1) U [Map] [REL] [

averageRating, U, (1, 1), U, [Double], [REL], Ø, ε, ε), numVotes, U, (1, 1), U, [Long], [REL], Ø, ε, ε)

PROPERTY-BASED INFERRER

- 1. Traversing collections by **attributes/properties** a. (PostgreSQL table - columns in rows)
- 2. For each property, a **property domain footprint** (PDF) is created
- 3. A structure-ignoring RSD is generated for each property (without recursion)
- 4. A schema of the collection is created based on the hierarchical names and PDFs

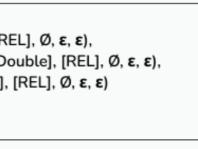
CANDIDATE MINER

- 1. Based on the **Property-based Inferrer**
- 2. We are looking for candidates for **unique identifiers**
- 3. We compare the generated PDFs to get **references**
- 4. We look for overlaps in values of size at least *k* to reveal **redundancies**
- 5. We get candidates for **Integrity Constraints**

PDFs of Kind Perso	on					PDFs of Kind Title					
hierarchical name	gDb.person/_/ nconst	gDb.person/_/ primaryName	gDb.person/_/ birthYear	gDb.person/_/ deathYear		hierarchical name	docDb.title/_/	docDb.title/_/directors/_/ nconst	docDb.title/_/directors/_/ primaryName	docDb.title/_/directors/_/ birthYear	docDb.title/_/directors/_/ deathYear
nique	Т	F	F	F		unique	Т	Т	F	F	F
quired	Т	Т	Т	F		required	Т	Т	Т	Т	F
iplicated	F	F	F	F		multiplicated	F	F	F	F	F
ntial	F	F	F	F		sequential	F	F	F	F	F
	99	216	14	112		min	12	99	216	14	98
	102	1290	2084	112		max	86	102	1290	2084	112
	100	574	704	112		average	49	100	574	704	105
Filter	[1, 1, 1, 0]	[1, 0, 1, 0]	[0, 1, 0, 1]	[1, 0, 0, 0]		bloomFilter	[0, 0, 0, 2]	[1, 1, 1, 0]	[1, 0, 1, 0]	[0, 1, 0, 1]	[1, 1, 0, 0]
	↑ identifier I reference						↑ identifier 	↑ identifier I			
					[redundancy		^	Ĵ	<u>}</u>	

2. For each record, a simple scheme (RSD) copying the structure (recursively) is

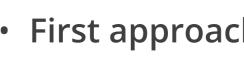
3. Schemas of individual records are **merged** into the overall description of the



ating RSDs? (_, U, (2, 2), U, [Map], [REL], (tconst, U, (2, 2), U, [String], [REL], Ø, ε, ε), (averageRating, U, (2, 2), U, [Double], [REL], Ø, ε, ε), (numVotes, U, (2, 2), U, [Long], [REL], Ø, ε, ε)]. ε. ε)

TABLE Rating								
tconst	averageRating		numVotes					
tt000001	5.7		1874					
tt000002	5.9		1874					
PDFs of Table Rating								
hierarchical name		rDb.rating/_/ tconst	rDb.rating/_/ averageRatir	ng rDb.rating/_/				
unique		Т	Т	F				
required		Т	Т	Т				
multiplicated		F	F	F				
sequential		F	F	F				
min		12	987	7				
max		86	7845	7				
average		49	4416	7				
bloomFilter		[0, 0, 0, 2]	[1, 0, 1, 0]	[0, 1, 0, 0]				

PDFs of Kind Title	



- - columnar, graph)
- between models
- Modular = easily extensible

- and redundancies)

			E)	XF	> E	
ht				D DDE data.c		
	5 000	_	Recor	rd Base	d Algor	ithm
	3 750					
lime [s]	2 500					
	1 250					
	0	1 k	2k	4k	8k	16k

- conferences:
- Technology

- 2022.61



FACULTY **OF MATHEMATICS** AND PHYSICS Charles University

CONTRIBUTIONS

• First approach for multi-model schema inference

• Process the data in an **agnostic way** of the underlying model • It can be used to generate data schemas in **commonly used** models (relational, document (XML, JSON), key-value,

• Finding interconnections between data within models and

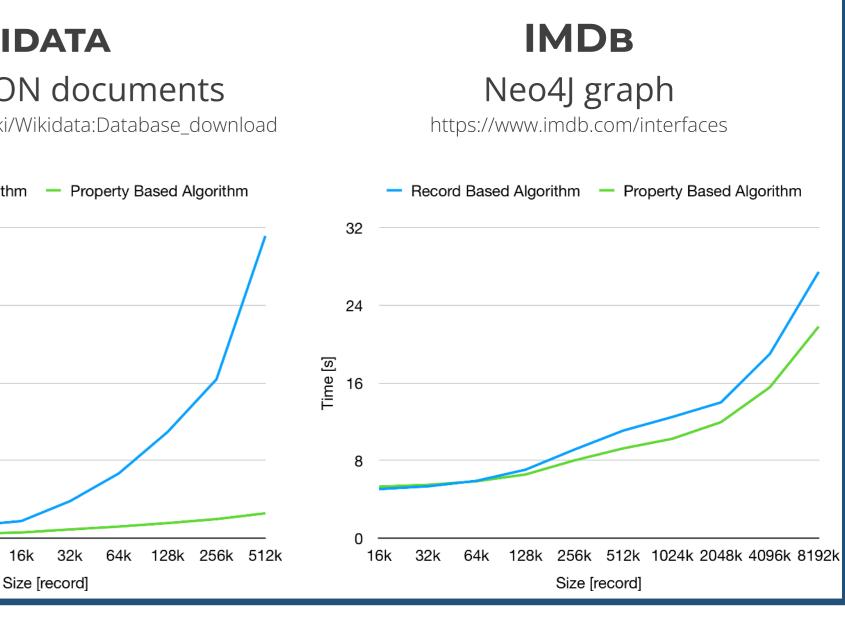
• Revealing **redundancy** in data

• Scalable, supporting distributed computing

• Extends schema generation possibilities for single-model data (uniqueness, regular expressions, more advanced references

• The proposed algorithms were **experimentally verified**

RIMENTAL EVALUATION



PUBLICATION ACTIVITY

• The results of the work were published (accepted) within

• EDBT: 25th International Conference on Extending Database

• ACM / IEEE 25th International Conference on Model Driven Engineering Languages and Systems (MODELS)

• Pavel Koupil, Sebastián Hricko, and Irena Holubová. Mm-infer: A tool for inference of multi-model schemas. In Julia Stoyanovich, Jens Teubner, Paolo Guagliardo, Milos Nikolic, Andreas Pieris, Jan Mühlig, Fatma Özcan, Sebastian Schelter, H. V. Jagadish, and Meihui Zhang, editors, Proceedings of the 25th International Conference on Extending Database Technology, EDBT 2022, Edinburgh, UK, March 29 - April 1, 2022, pages 2:566–2:569. OpenProceedings.org, 2022

 Pavel Koupil, Sebastián Hricko, and Irena Holubová. A universal approach for multi-model schema inference. J. Big Data, 9(1):97,