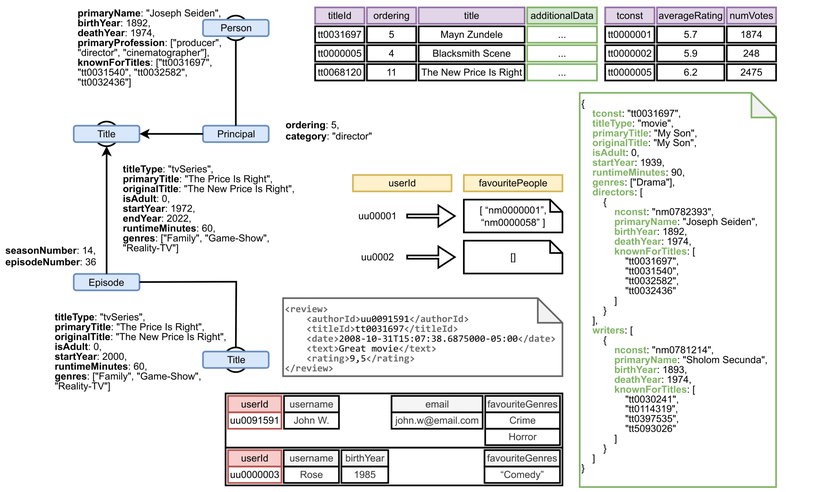
Schema Inference for Multi-model Data

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

FACULTY OF MATHEMATICS AND PHYSICS Charles University

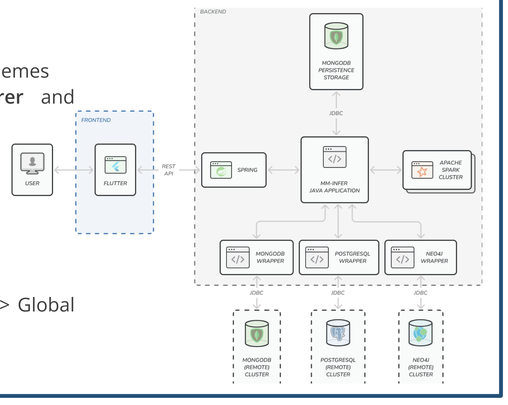
Problem Definition



- 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 model?
  - How is the data **interconnected** between models?
- We want to obtain a **global view** of the data model and identify **references** and **redundancies** in the dataset

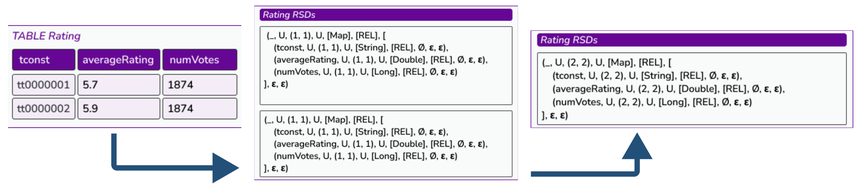
Proposed Approach

1. **Unification** of models
2. **Analysis and reduction** of unified schemes
   a. Algorithms **Record-based Inferrer** and **Property-based Inferrer**
   b. Apache Spark MapReduce
3. **Discovering** candidates
   a. IO, references, redundancies
   b. **Candidate Miner** algorithm
   c. Heuristics
4. **Evaluation** of candidates
5. **Connecting** reduced local schemes -> Global scheme



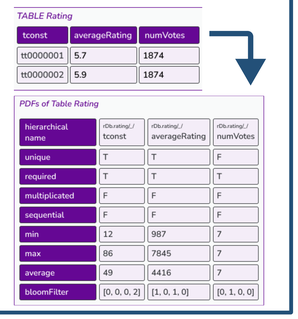
Record-based Inferrer

1. Traversing collections by **records** (i.e. PostgreSQL table - rows)
2. For each record, a **simple scheme (RSD)** copying the structure (recursively) is created
3. Schemas of individual records are **merged** into the overall description of the collection



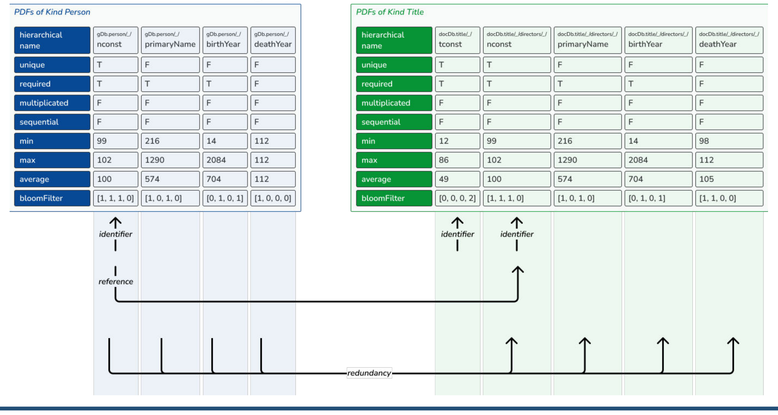
Property-based Inferrer

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**



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, columnar, graph)
  - Finding **interconnections** between data within models and between models
  - Revealing **redundancy** in data
- **Modular** = easily extensible
- **Scalable**, supporting distributed computing
- Extends schema generation possibilities for **single-model** data (uniqueness, regular expressions, more advanced references and redundancies)
- The proposed algorithms were **experimentally verified**

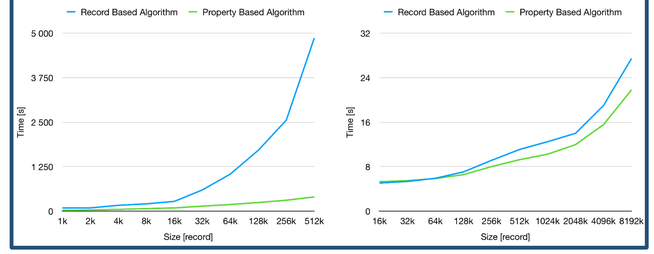
Experimental Evaluation

Wikidata

MongoDB JSON documents
https://www.wikidata.org/wiki/Wikidata:Database_download

IMDb

Neo4J graph
https://www.imdb.com/interfaces



Publication Activity

- The results of the work were published (accepted) within conferences:
  - **EDBT: 25th International Conference on Extending Database Technology**
  - **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, 2022.61