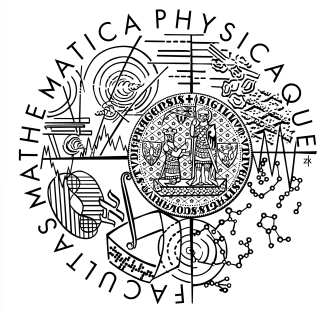
Katedra softwarového inženýrství



MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

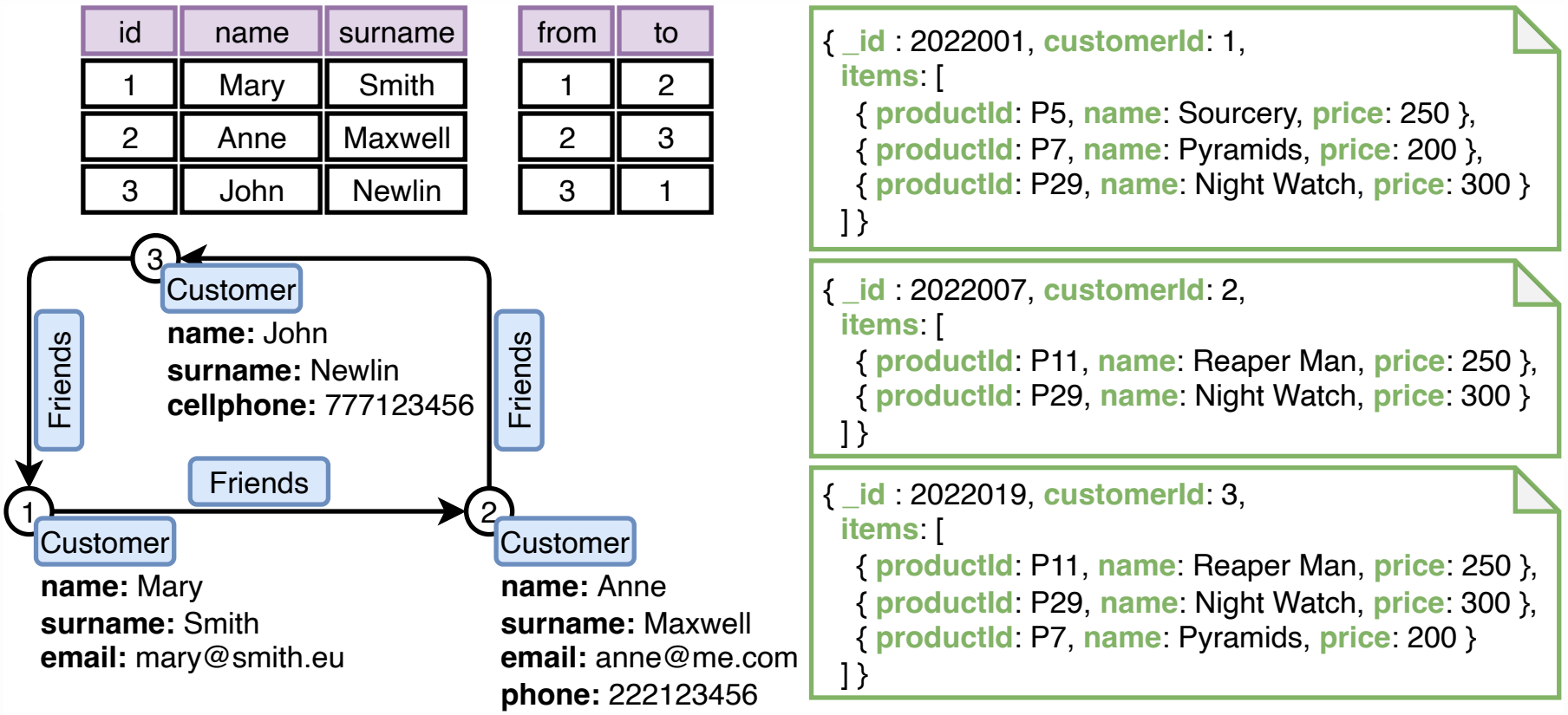
Bc. Daniel Crha

Unifikované dotazování nad multi-modelovými daty

Obhajoba diplomové práce

8. února 2023

Multi-modelová data



Obrázek: Ukázka multi-modelových dat – data relační, grafová a dokumentová.

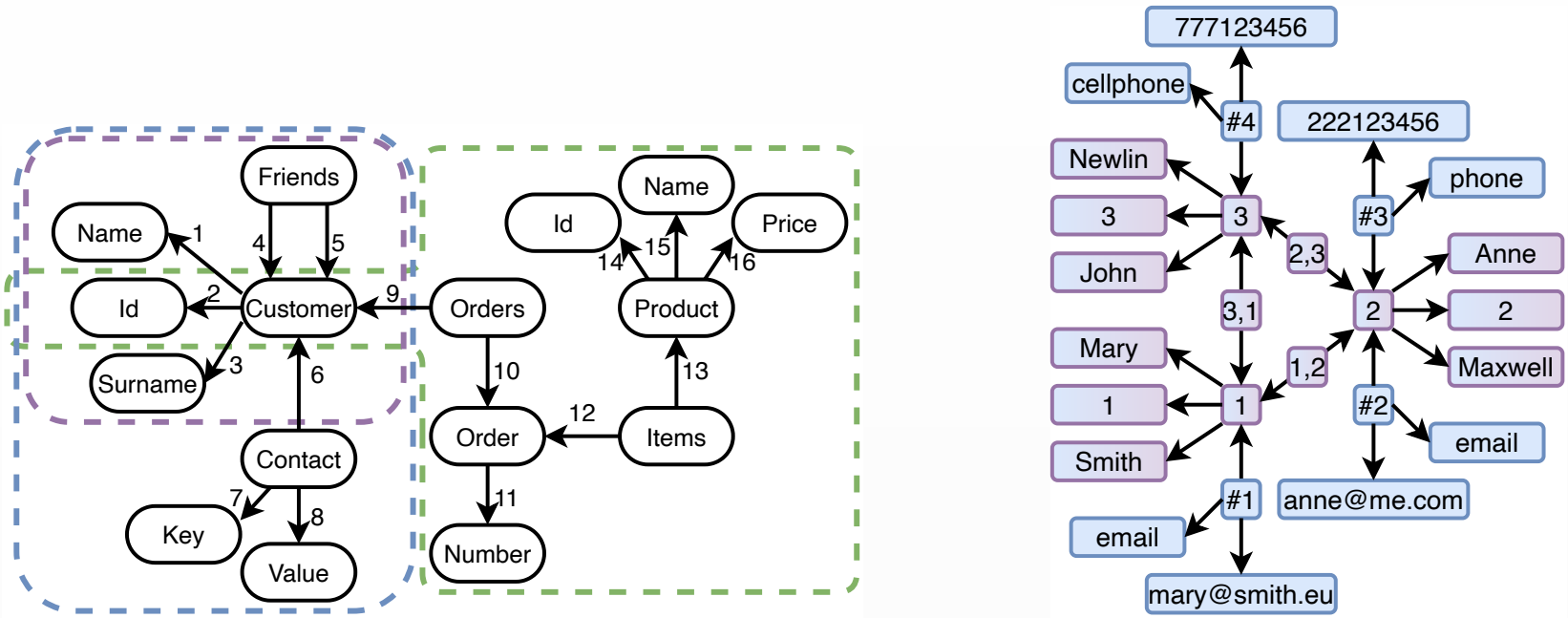
Úvod

- Co jsou multi-modelová data?
  - Motivace: rychlost, přirozenost reprezentace, redundance
- Dotazovat lze pouze jednotlivé modely zvlášť
- Chceme data dotazovat unifikovaně
- Jedno řešení: MultiCategory: Multi-model Query Processing Meets Category Theory and Functional Programming (Valter Uotila et al., 2021)
  - Funkcionální programování (Haskell)
  - Neuvažuje datovou redundanci
  - Náročné na použití pro širší spektrum uživatelů

Předpoklady

- Cíl: unifikované dotazování dat ve více databázích
  - Jednotný dotazovací jazyk
  - Chceme ignorovat specifika jednotlivých datových modelů
- Potřebujeme unifikovanou abstraktní reprezentaci dat
  - Reprezentace založená na teorii kategorií
  - Unifikuje specifika populárních modelů – relační, dokumentový, grafový, key-value, sloupcový
  - Budeme nad touto reprezentaci dotazovat
- Tři zásadní koncepty – schématická kategorie, instanční kategorie a mapování dat z nativní do kategorické reprezentace

Kategorický model



Obrázek: Schématická kategorie (vlevo) a instanční kategorie (vpravo).

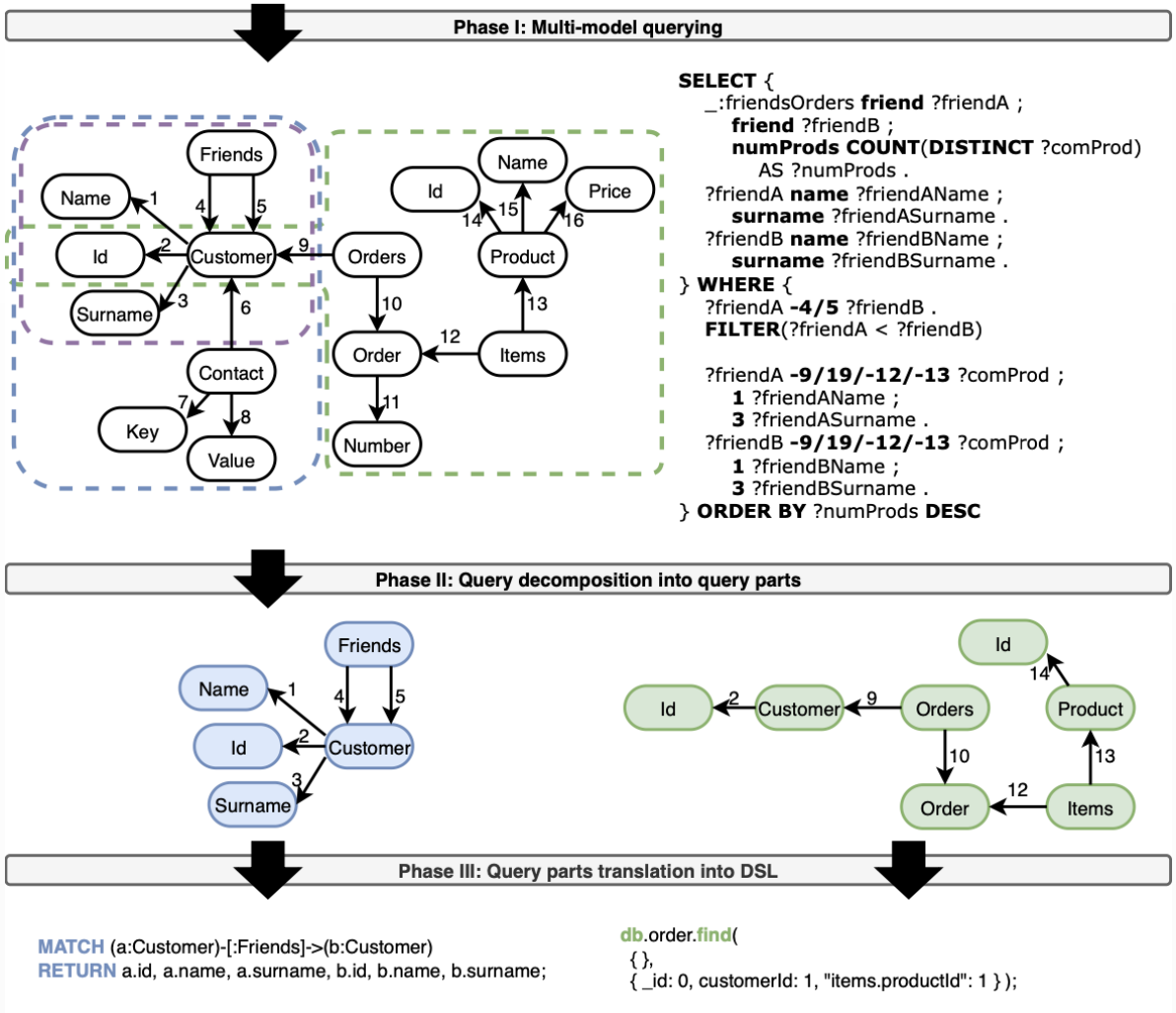
Návrh přístupu

- Máme unifikovanou kategorickou reprezentaci, potřebujeme kategorický dotazovací jazyk
- Kategorii lze reprezentovat jako multigraf
  - Můžeme použít grafový dotazovací jazyk jako základ
  - Jazyk SPARQL - expresivita, známá technologie

Přínos 1: Jazyk MMQL

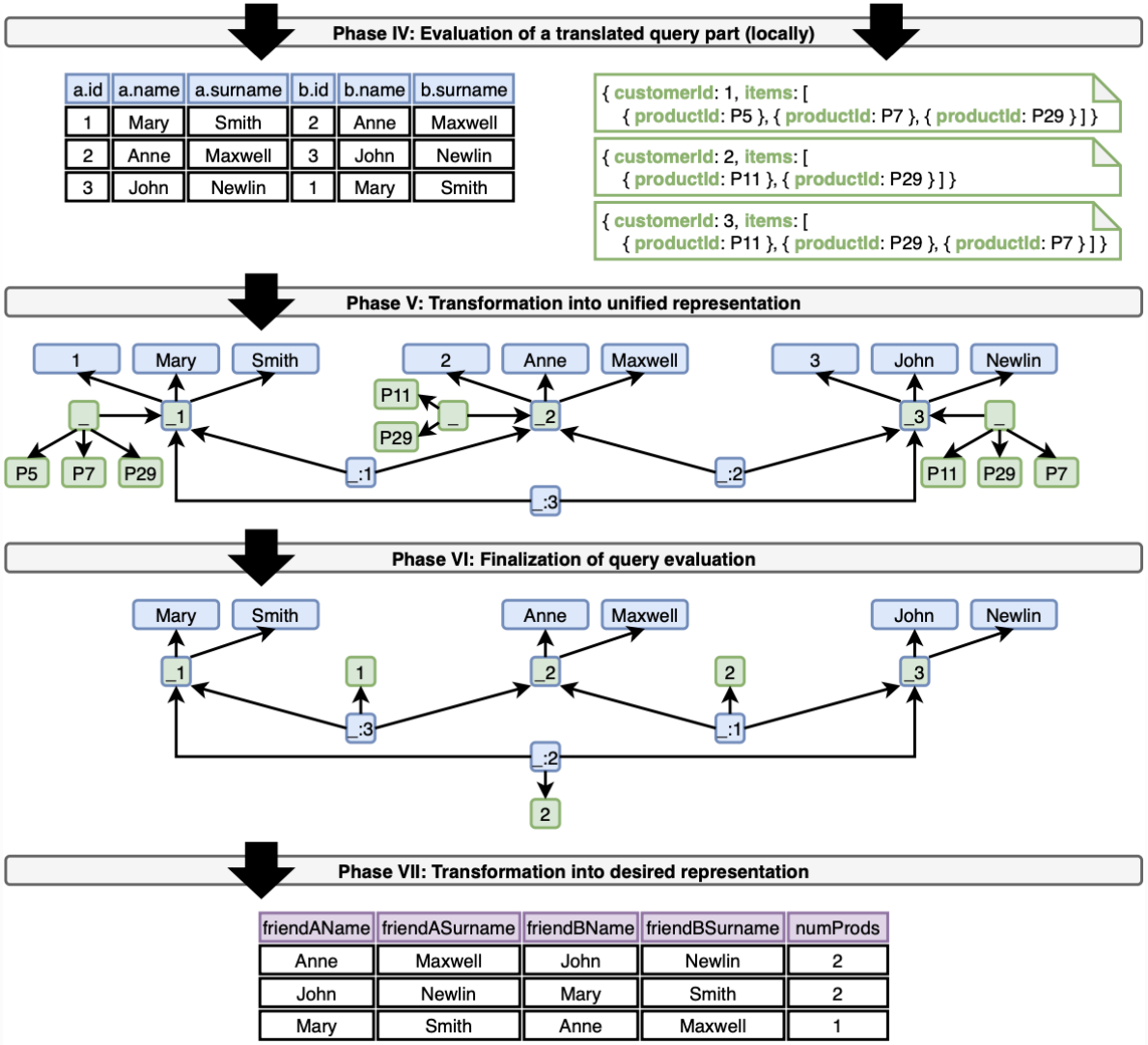
- Dotazovací jazyk Multi-Model Query Language
  - Analýza existujících grafových dotazovacích jazyků
  - Inspirace SPARQL – kategorická doména, objekty a morfismy
  - Dotazujeme schématickou kategorii, výsledkem je instanční kategorie
  - Matchování grafových vzorů
  - Projekce, selekce, filtrování, agregace, řazení
  - Množinové operace, vnořené dotazy, více proměnných pro jeden schématický objekt
- Výstupem je i gramatika MMQL

Přínos 2: Dotazovací algoritmy



Obrázek: Kroky 1-3 zpracování MMQL dotazu.

Přínos 2: Dotazovací algoritmy



Obrázek: Kroky 4-7 zpracování MMQL dotazu.

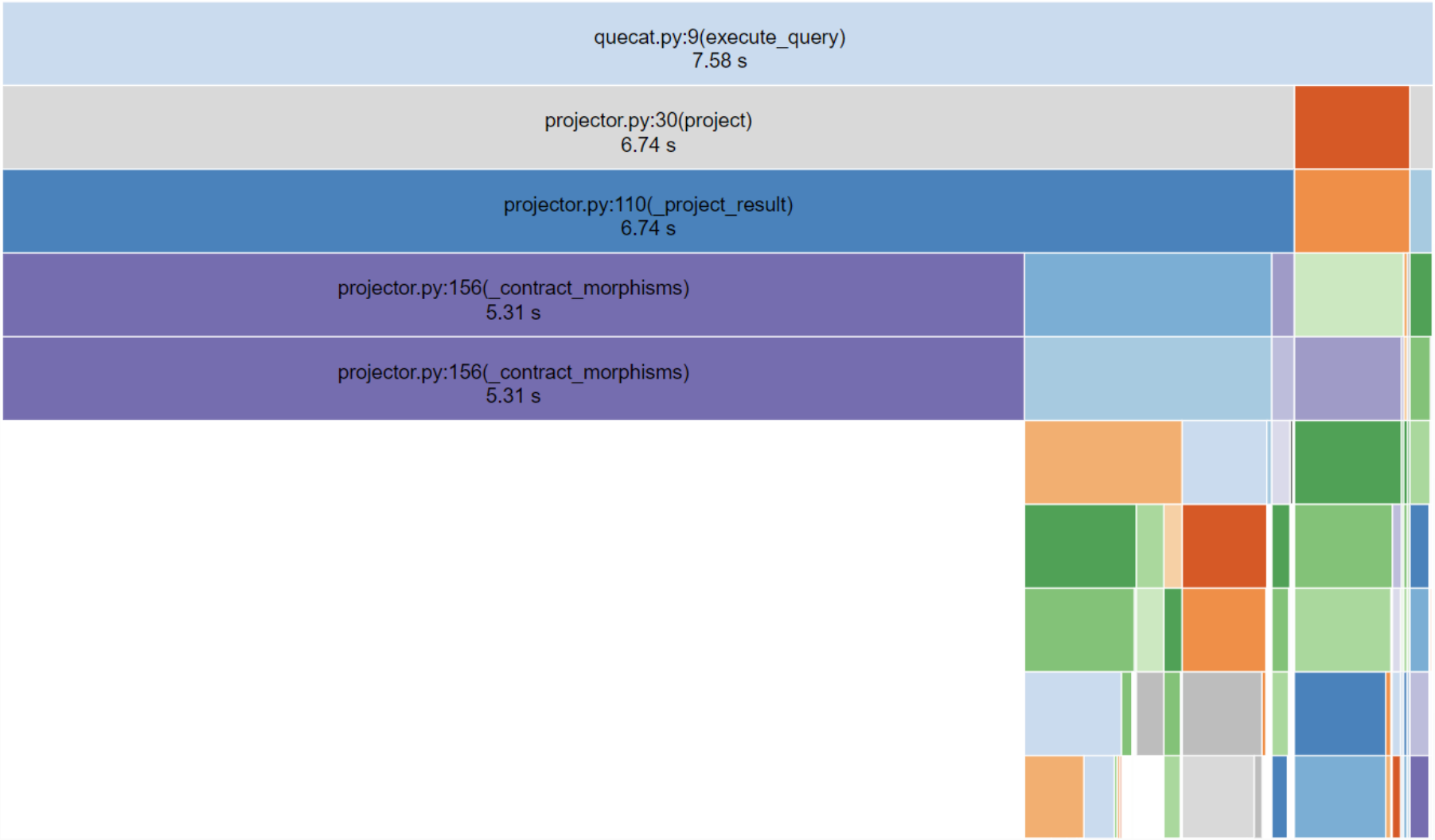
Přínos 2: Dotazovací algoritmy

- Dotazovací jazyk je nutné implementovat
  - Velice netriviální
  - Pro multi-modelovou doménu není dostatek relevantních zdrojů
- Návrh přístupu a specifikace algoritmů pro implementaci MMQL
  1. Tvorba plánů dotazu
  2. Tvorba plánů spojení
  3. Výběr nejlepšího plánu dotazu
  4. Překlad query parts do nativních dotazů
  5. Spojení výsledků
  6. Projekce a další operace
  7. Transformace do finální reprezentace
- Zaměření na jednoduchost a pochopitelnost, výkon je pro nás vedlejší

Přínos 3: Nástroj MM-quecat

- Ověření validity návrhu
- Nástroj MM-quecat
  - Implementace konceptů MMQL
  - Podporuje databáze PostgreSQL, MongoDB
  - Selekce, projekce, spojení v rámci databáze i mezi databázemi
- Experimentální evaluace s profilováním
  - Odhalení neefektivních částí dotazovacího algoritmu
  - Efektivita není primárním cílem práce

Experimentální evaluace MM-quecat

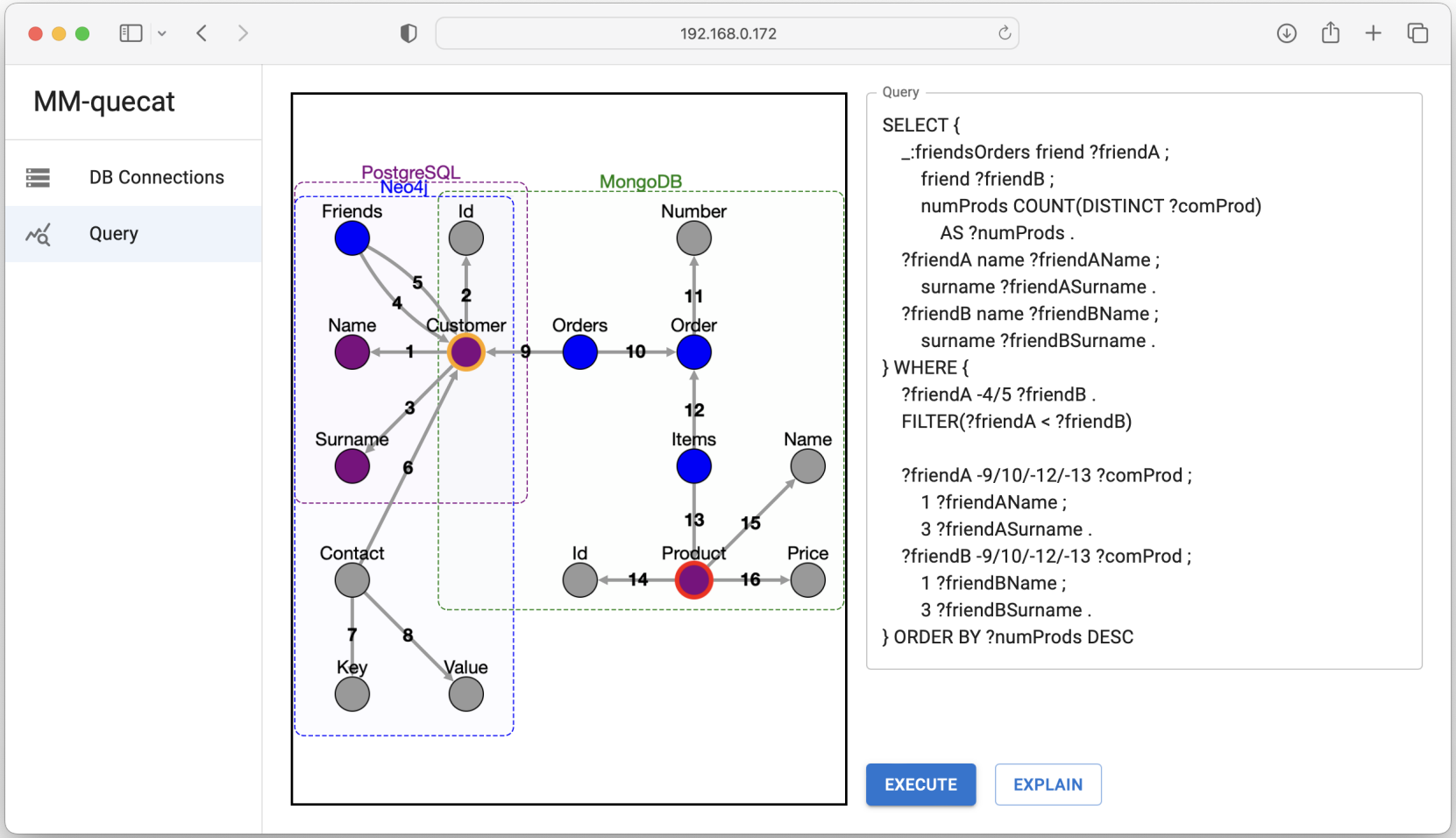


Obrázek: Vizualizace výstupu profileru pro experiment s MongoDB.

Publikace

- Demo článek MM-quecat: A Tool for Unified Querying of Multi-Model Data
  - Autoři: Pavel Koupil, Daniel Crha, Irena Holubová
  - Přijat na konferenci EDBT 2023 (CORE A)
  - Rozšířená verze nástroje z diplomové práce
- Příprava rozšířeného žurnálového článku o MMQL

Uživatelské rozhraní MM-quecat



Obrázek: Návrh uživatelského rozhraní MM-quecat.

Závěr

- Hlavní přínos práce: jazyk MMQL
  - Jednoduché dotazy jsou jednoduché, složité jsou možné
  - Podpůrné algoritmy pro implementaci jazyka
  - Unifikované dotazování multi-modelových dat
- Proof-of-concept implementace MM-quecat
  - Verifikace návrhu
- Kritická evaluace návrhu
  - Identifikace neefektivních částí
  - Návrhy na další zlepšení a rozšíření
- První krok v unifikovaném dotazování multi-modelových dat, poukázání na další zajímavé problémy

Děkuji za pozornost!

Za ochotu a čas mně věnovaný při vypracování diplomové práce děkuji též své vedoucí práce doc. RNDr. Ireně Holubové, Ph.D.

Připomínky oponenta – Čitelnost MMQL

- “s. 19: Požadavky na kategoriální dotazovací jazyk zmiňují čitelnost jazyka. Použití signatur v jazyku MMQL (viz kap. 4) ovšem jeho čitelnost nikterak nezvyšuje, spíše naopak.”
  - MMQL počítá s alfanumerickými signaturami
  - Čísla jsou automaticky vygenerovaná z MM-evocat, lze používat i lepší identifikátory

Připomínky oponenta – Mappings

- “s. 42: krok 4 – jak systém pro vyhodnocení dotazu rozezná, zda má použít při přístupu k objektu pole nebo vnořený objekt. Jde o 2 různé dotazy.”
  - Koncept mappings – podkapitola 2.5
  - Mapping definuje mapování nativní reprezentace do kategorické reprezentace

Připomínky oponenta – ER Diagramy

- Používám zjednodušené diagramy pro přehlednost
  - Chybějící kardinality dle kontextu
  - Klíčové atributy dle kořenu kindu
  - _src a _tgt atributy udávají směr hrany
  - Diagramy obsahují mapy klíč/hodnota
- To platí i pro schématické kategorie

Připomínky oponenta – Plány dotazu

- “s. 88: který typ uživatele bude analyzovat různé plány dotazu?”
  - Cena plánu se může měnit (přidání indexu, velká zátěž na konkrétní databázi, ...)
  - Můžeme chtít přidat indexy podle daného plánu