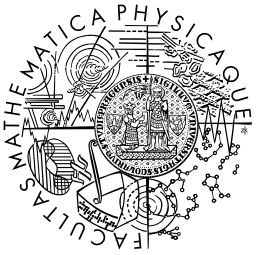

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Richard Eliáš

Analýza datových toků ve databázových systémech

Katedra distribuovaných a spolehlivých systémů

Vedoucí diplomové práce:	RNDr. Pavel Parízek, Ph.D.
Studijní program:	Informatika
Studijní obor:	Umělá inteligence

Praha 2019

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů (dále jen "autorský zákon"), především § 35 a § 60 autorského zákona upravující školní dílo. V případě počítačových programů, jež jsou součástí mojí práce či její přílohou, a veškeré související dokumentace k počítačovým programům (dále jen "software"), uděluji tzv. MIT Licenci.

Obsahem MIT Licence je oprávnění bezúplatně užít software. Toto oprávnění uděluji každé osobě, která má o užití software zájem.

Každá osoba je oprávněna pořídit si kopii software (včetně související dokumentace) bez jakéhokoli omezení a dále je oprávněna bez jakéhokoli omezení software zejména užívat, kopírovat, upravovat, sloučit, publikovat, distribuovat, poskytovat podlicence a / nebo prodávat kopie software a umožnit výkon těchto práv osobám, kterým bude dále software poskytnut. Způsoby užití software ani rozsahu tohoto užití nejsou jakkoli omezeny.

Osoba, která má o užití software zájem je povinna připojit text licenčních podmínek následujícího znění:

Copyright © 2019 Richard Eliáš

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

V ..... dne ..... Podpis autora

Název práce: Analýza datových toků ve databázových systémech

Autor: Richard Eliáš

Katedra: Katedra distribuovaných a spolehlivých systémů

Vedoucí bakalářské práce: RNDr. Pavel Parízek, Ph.D., Katedra distribuovaných a spolehlivých systémů

Abstrakt: TODO Abstrakt

Klíčová slova: TODO Klucove slova

Title: Analyzing Data Lineage in Database Frameworks

Author: Richard Eliáš

Department: Department of Distributed and Dependable Systems

Supervisor: RNDr. Pavel Parízek, Ph.D., Department of Distributed and Dependable Systems

Abstract: TODO Abstrakt

Keywords: TODO Klucove slova

TODO Podakovanie

Obsah

Úvod

TODO Nejaky uvod...

1. Úvod do statickej analýzy

1.1 Motivácia

TODO motivacia - naco je to cele dobre

1.2 MANTA

1.2.1 História

História nástroja Manta Flow siaha do roku 2008, kedy ho vyvinula česká konzultačná spoločnosť Profinit s.r.o. ako interný nástroj. Následne v roku 2013 sa autori projektu rozhodli založiť samostatnú spoločnosť a pokračovať v jeho vývoji. V čele s Ing. Tomášom Krátkým založili spoločnosť Manta Tools s.r.o. známu ako MANTA.

Spoločnosť sa spolu s ČVUT zapojila do krantových programov ALFA a EPSILON Technologickej Agentury Českej Republiky (TAČR). Vďaka nim získala v rokoch 2013 a 2017 dva granty v celkovej výške 1,49 milióna dolárov.

Spoločnosť sa presadila na zahraničnom trhu predovšetkým po výťazstve v súťaži "Czech ICT Incubator @ Silicon Valley" v roku 2014, ktorú usporadúva Czech ICT Alliance. Po víťazstve bola založená prvá americká pobočka v San Francisku.

V súčastnosti pôsobí MANTA celosvetovo prostredníctvom vlastných pobočiek a siete regionálnych partnerov. Medzi zákazníkov patrí napríklad spoločnosť Paypal, OBI, Vodafone, alebo Comcast.

O spoločnosti sa môžeme viac dočítať v článkoch od autorov Sedlák (2015) a Pavlunová (2017).

1.2.2 Manta Flow

Manta Flow je nástrok umožňujúci automatickú analýzu programovacieho kódu (SQL, Java) a následný popis transformačnej logiky, ktorý v ňom je obsiahnutý. Software je schopný rozpoznať aj ťažko čitateľné, na mieru napísané riadky programovacieho kódu. Vďaka tejto vlastnosti dokáže v pomerne krátkom čase (obvykle niekoľko hodín) automaticky prečítať databáze o rozsahu stotisícov i niekoľkých miliónov údajov a zostaviť z nich prehľadnú mapu datových tokov naprieč BI prostredím (Data Lineage). To sa v praxi využíva najmä k optimalizácií datových skladov, znižovaniu nákladov na vývoj softwaru, vykonávanie dopadových analýz a pri dokumentovaní prostredí pre potreby regulačných úradov.

2. Prehľad používaných frameworkov na komunikáciu s databázou

V tejto kapitole popíšeme techniky a niektoré frameworky, ktoré umožňujú aplikácii prístup k rôznym druhom databáz.

Prácu s frameworkami si budeme ilustrovať na získaní dát pre triedu Category 2.1 z databáze.

```
1  public class Category {
2      private Integer id;
3      private String name;
4      private String description;
5
6      public Integer getId() {
7          return id;
8      }
9
10     public void setId(Integer id) {
11         this.id = id;
12     }
13
14     public String getName() {
15         return name;
16     }
17
18     public void setName(String name) {
19         this.name = name;
20     }
21
22     public String getDescription() {
23         return description;
24     }
25
26     public void setDescription(String description) {
27         this.description = description;
28     }
29 }
```

Listing 2.1: Príklad objektu získavaného z databáze

2.1 JDBC

Prístup aplikácie k databáze je v Jave štandardne cez Java Database Connectivity (JDBC) API. Rozhranie je implementované pre rôzne typy databáz. Dovoľuje pristupovať k dátam priamo s použitím Javy. Architektúra pre prístup k databáze je popísaná v článku JDBC Overview.

Komunikácia s databázou sa uskutočňuje pomocou rozhraní a tried z balíka java.sql. Niektoré zo základných štruktúr si dôkladnejšie popíšeme.

- Connection - objekt zabezpečujúci pripojenie do databáze
- Statement, PreparedStatement, CallableStatement - objekty obaľujúce jednotlivé volania databáze

- ResultSet - objekt obaľujúci výsledok volaní databáze

Na príklade 2.2 si ukážeme načítanie triedy Category z Oracle databáze. Riadok 3 zaregistruje Oracle ovládač databáze. Následne riadok 7 vytvorí pripojenie k databáze, riadok 14 vykoná sql príkaz a na ďalších riadkach sa z výsledku získajú hodnoty zo stĺpcov ID, DESCRIPTION a NAME.

```
1  public Category getCategoryById(int categoryId)
2          throws ClassNotFoundException, SQLException {
3    Class.forName("oracle.jdbc.driver.OracleDriver");
4    Connection connection = null;
5    try {
6      String url = "jdbc:oracle:driver:user/password@database";
7      Connection = DriverManager.getConnection(url);
8      PreparedStatement statement = null;
9      try {
10       String sql = "SELECT ID, DESC, NAME FROM CATEGORY_TABLE WHERE ID = ?";
11       statement = connection.prepareStatement(sql);
12       statement.setInt(1, categoryId);
13
14       final ResultSet rs = statement.executeQuery();
15       final Category category = new Category();
16
17       category.setId(rs.getInt("ID"));
18       category.setDescription(rs.getString("DESC"));
19       category.setName(rs.getString("NAME"));
20
21       return category;
22     } finally {
23       if (statement != null) {
24         statement.close();
25       }
26     }
27   } finally {
28     if (connection != null) {
29       connection.close();
30     }
31 }
```

Listing 2.2: Príklad použitia JDBC

TODO - popisovat na prikladoch?

V štandardnej edícií javy sa objavuje aj balík java.sql, ktorý poskytuje jednoduchšie API pre konfiguráciu databáze. Rovnako pridáva funkcie pre distribuované tranzakcie a connection pooling.

2.2 Spring JDBC Framework

Spring JDBC Framework je istou nadstavbou nad klasickým JDBC API, ktorého cieľom je uľahčiť jeho používanie. Uľahčuje ošetrovanie výnimiek pri komunikácií s databázou, zjednodušuje možnosti používania tranzakcií, znižuje množstvo opakujúceho sa kódu pri získavaní výsledkov z databázových volaní.

TODO - popisovat na prikladoch?

2.3 MyBatis

MyBatis

2.4 Hibernate

2.5 Spark

2.6 Kafka

3. Ďalšie možnosti pre vstup/výstup

3.1 Súbory, sockety, ...

4. Statická analýza

4.1 WALA framework

4.2 Nadstavba analýzy

5. Implementácia

5.1 Rozhranie

5.2 Jdbc

5.3 MyBatis

5.4 JdbcTemplate

5.5 Kafka

6. Výsledky

Záver

TODO Nejaky zaver

Zoznam použitej literatúry

java.sql. Dokumentácia pre balík java.sql. https://docs.oracle.com/javase/8/docs/api/java/sql/package-summary.html. Dostupné 28.3.2018.

JDBC Overview. JDBC Overview. https://www.oracle.com/technetwork/java/overview-141217.html. Dostupné 28.3.2018.

MyBatis. Manuál pre MyBatis Framework. http://www.mybatis.org/mybatis-3/. Dostupné 28.3.2018.

Pavlunová, A. (2017). Akademik a manager Tomáš Krátký: Čeští studenti přemýšlí v souvislostech, firmy jim ale moc nepomáhají. http://www.czechcrunch.cz/2017/06/akademik-a-manager-tomas-kratky-cesti-studenti-premysli-v-souvislostech-firmy-jim-ale-moc Dostupné 28.3.2018.

Sedlák, J. (2015). V Dejvicích roste další velká česká softwarová věc. https://connect.zive.cz/clanky/v-dejvicich-roste-dalsi-velka-ceska-softwarova-vec/sc-320-a-177282. Dostupné 28.3.2018.

Spring JDBC. Manuál pre Spring JDBC Framework. https://docs.spring.io/spring/docs/2.0.x/reference/jdbc.html. Dostupné 28.3.2018.

Zoznam obrázkov