Slovenská technická univerzita
Fakulta informatiky a informačných technológii
Ilkovičova 3, 812 19 Bratislava

Databázové systémy

Analytický report

Matúš Cuper

73688

Cvičiaci: Ing. Juraj Žoffčák
Študijný odbor: Informatika
Ročník: 3. Bc
Akademický rok: 2016/2017

Obsah

1. Zadanie

Implementovaný relačný model musí byť naplnený dostatočným množstvom údajov. Požaduje sa stručný opis spôsobu, akým boli dáta vygenerované a ich počty riadkov. Každý numerický atribút bude opísaný sumarizáciou zahŕňajúcou minimum, maximum, medián, dolný a horný kvartil. Nominálne atribúty budú obahovať početnosti jednotlivých hodnôt.

2. Generovanie dát

Dáta som generoval a vkladal do databázy pomocou programu, ktorý nie je súčasťou odovzdania, nakoľko samotný generátor je open source projekt a práca s databázou sa nachádza aj v odovzdávanom programe.

Dáta boli vygenerované pomocou generátoru dát jFairy (https://github.com/Codearte/jfairy). Výstupom sú generované údaje o osobách a inštitúciách, v ktorých pracujú. Údaje sa generujú na základe súboru v YAML formáte, pre každý jazyk osobitne. Slovenčina samozrejme zastúpenie nemá, a preto som musel dáta doplniť a prekompilovať program. Mená, priezviská adresy, ktoré som použil sa nachádzajú v adresári *data*. Z poskytovanej funkcionality som použil iba generovanie osôb a adries.

Následne som vo vytvorenej databáze vytvoril tabuľky a najskôr naplnil číselníkové tabuľky, ktorých dáta sa taktiež nachádzajú v adresári *data*. Potom som vložil študentov, ktorý sa generovali volaním generátora. Keďže pri vkladaní týchto dát sa primárne kľúče automaticky generovali a žiadne dáta som nemazal, mohol som zvoliť náhodné číslo z rozsahu primárnych kľúčov a použiť ho ako cudzí kľúč do inej tabuľky. Takto vznikali spojovacie tabuľky, kde vygenerovaný dátum je z rozsahu študentovho narodenia až po dnešný dátum.

2.1. Tabuľka statuses

Ide o číselník, ktorého obsah je zhodný s rovnomenným súborom v adresári *data*. Obsahuje statusy jednotlivých prihlášok žiakov.

2.2. Tabuľka fields_of_study

Ide o číselník, ktorého obsah je zhodný s rovnomenným súborom v adresári *data*. Obsahuje názvy študijných odborov, ktoré je možné na Slovensku študovať na vysokých školách a univertizách.

2.3. Tabuľka award_levels

Ide o číselník, ktorého obsah je zhodný s rovnomenným súborom v adresári *data*. Opisuje úroveň, na akej študent vyhral danú súťaž, čiže napr. krajské kolo alebo celoštátne.

2.4. Tabuľka award_names

Ide o číselník, ktorého obsah je zhodný s rovnomenným súborom v adresári *data*. Obsahuje názvy súťaží, väčšina sú v oblasti publikovania, ale aj zopár hackathonov.

2.5. Tabuľka subjects

Ide o číselník, ktorého obsah je zhodný s rovnomenným súborom v adresári *data*. Obsahuje predmety, ktoré sa všeobecne nachádzajú na stredných školách a študenti z nich maturujú.

2.6. Tabuľka universities

Ide o číselník, ktorého obsah je zhodný s rovnomenným súborom v adresári *data*. Obsahuje všetky vysoké školy, ktoré sa nachádzajú na Slovensku. Adresa je získaná z vygenerovaného študenta, adresa sa vloží do tabuľky a študent sa zahodí.

2.7. Tabuľka secondary_schools

Ide o číselník, ktorého obsah je zhodný s rovnomenným súborom v adresári *data*. Obsahuje niektoré stredné školy na Slovensku. Adresa je opäť získaná vygenerovaním študenta, z ktorého je použitá iba adresa.

2.8. Tabuľka fos_at_universities

Ide o spojovaciu tabuľku, ktorá spája univerzity a študijné odbory. Vygenerujú sa cudzie kľúče a pred vložením do tabuľky sa overí, že sa tam táto kombinácia ešte nenachádza. Po dosiahnutí požadovaného počtu kombinácií, generovanie skončí.

2.9. Tabuľka students

Študent je vygenerovaný pomocou jFairy a stredná škola, na ktorej študuje alebo maturoval mu je vygenerovaná náhodne a je vložená ako cudzí kľúč do tejto tabuľky. Osobné údaje študenta sú taktiež generované pomocou jFairy. Po dosiahnutí požadovaného počtu záznamov, generovanie skončí.

2.10. Tabuľka graduations_from_ss

Tabuľka reprezentuje maturitnú skúšku študenta na jeho strednej školy z predmetu, na ktorý odkazuje pomocou cudzieho kľúča. Vygenerovaný dátum je po dátume narodenia študenta a známka z maturity je v rozsahu od 1 do 4. Po dosiahnutí požadovaného počtu maturít, generovanie skončí.

2.11. Tabuľka awards

Tabuľka reprezentuje vyznamenania, ktoré študent počas svojho života získal. Každé ocenenie má svoj názov a úroveň, na ktorej sa koná. Úrovňou možno rozumieť geografické územie, ktoré vyznamenanie pokrýva. Po dosiahnutí požadovaného počtu ocenení, generovanie skončí.

2.12. Tabuľka graduations

Tabuľka reprezentuje ukončenie štúdia na vysokej škole. To môže byť úspešné alebo neúspešné, čo je zachytené pomocou atribútu *graduated*. Vygenerované sú cudzie kľúče do tabuľky študentvo a tabuľky študijných odborov na univerzitách. Vygenerovaný dátum je neskôr ako dátum narodenia a začiatok pôsobenia na univerzite je logicky skôr ako jeho ukončenie. Žiadne ďalšie pravidlo sa neuplatňuje pri generovaní dát, čiže môže vzniknúť študent v predškolskom veku, ktorý vyštudoval viacero vysokých škôl behom pár dní či mesiacov. Po dosiahnutí požadovaného počtu absolventov, generovanie skončí.

2.13. Tabuľka registrations

Tabuľka reprezentuje prihlášky, ktoré boli podané študentami na nejaký študijný odbor. Každá prihláška má nejaký stav, v ktorom sa nachádza, ktorý platí od nejakého dátumu až po súčasnosť. Vygenerovaný dátum je po dátume narodenia študenta. Stav prihlášky je pridelený náhodne. V tabuľke tak nie sú zachytené postupnosti pri spracovaní prihlášky. Po dosiahnutí požadovaného počtu prihlášok, generovanie skončí.

3. Veľkosti tabuliek

Nasledujúca tabuľka zobrazuje počet riadkov v jednotlivých tabuľkách.

Názov tabuľky	**Počet riadkov**
statuses	6
fields_of_study	328
award_levels	5
award_names	87
subjects	21
universities	35
secondary_schools	71
fos_at_universities	7 000
students	1 000 000
graduations_from_ss	1 250 000
awards	500 000
graduations	800 000
registrations	200 000

4. Početnosti nominálnych atribútov

Nasledujúce tabuľky zobrazujú početnosti jednotlivých nominálnych atribútov

Počet registrácií, ktoré sa nachádzajú v jednotlivých stavoch

statuses.name	**Počet riadkov**
spracováva sa	33 762
zamietnutá	33 550
zrušená	33 279

čaká na platbu	33 264
podaná	33 160
schválená	32 985

Histogram študijných odborov nachádzajúcich sa na rovnakom počte univerzít. Prvý riadok znamená, že v databáze je evidovaných 14 študijných odborov, ktoré sa nachádzajú rovnako na 4 univerzitách

Počet študijných odborov na univerzitách	**Počet riadkov**
14	4
15	5
16	8
17	22
18	22
19	33
20	34
21	35
22	37
23	44
24	32
25	27
26	17
27	6
28	2

Počet ocenení udelených na jednotlivých úrovniach

award_levels.name	**Počet riadkov**
celoštátne kolo	100 317
krajské kolo	99 963
európske majstrovstvá	99 953
štátna súťaž	99 922

okresné kolo	99 845

Počty študentov, ktorí boli ocenený v rovnakej súťaži. Iba prvých 10 súťaží

award_names.name	Počet riadkov
The New School Competition	5 937
Stockholm Innovation Scholarship Competition	5 926
Short Story Competition	5 925
Fordham Business Challenge	5 896
Inkitt Novel Competition	5 862
Short Story Award	5 854
New Voices Competition	5 849
Bath Short Story Award	5 849
David Nathan Meyerson	5 847
Short Fiction Prize	5 843

Prvých 10 predmetov na stredných školách, z ktorých sa najviac maturovalo.

subjects.name	Počet riadkov
Ruský jazyk	59 922
Logika	59 874
Slovenský jazyk a literatúra	59 868
Nemecký jazyk	59 804
Informatika	59 777
Latinský jazyk	59 763
Občianska náuka	59 581
Fyzika	59 574
Španielsky jazyk	59 557
Dejepis	59 539

Počet študijných odborov na jednotlivých vysokých školách, iba prvých 10 s najväčším počtom študjiných odborov

universities.name	Počet riadkov
Univerzita Pavla Jozefa Šafárika v Košiciach	217
Akadémia médií odborná vysoká škola mediálnej a marketingovej komunikácie v Bratislave	215
Vysoká škola zdravotníctva a sociálnej práce sv. Alžbety v Bratislave	213
Bratislavská medzinárodná škola liberálnych štúdií	213
Vysoká škola ekonómie a manažmentu verejnej správy v Bratislave	211
Technická univerzita vo Zvolene	210
Vysoká škola múzických umení v Bratislave	210
Slovenská technická univerzita v Bratislave	208
Trnavská univerzita v Trnave	208
Slovenská poľnohospodárska univerzita v Nitre	207

Počet študentov, ktorý chodili na tie najpopulárnejšie stredné školy

secondary_schools.name	Počet riadkov
Gymnázium Andreja Kmeťa	14 335
Súkromná stredná odborná škola ochrany osôb a majetku	14 313

Stredná odborná škola techniky a služieb	14 298
Stredná odborná škola sklárska Lednické Rovne	14 292
Stredná odborná škola drevárska	14 278
Pedagogická a kultúrna akadémia	14 245
Stredná priemyselná škola strojnícka	14 221
Gymnázium Ivana Horvátha	14 219
Súkromná stredná odborná škola SD Jednota	14 210
Katolícke gymnázium Štefana Moysesa Banská Bystrica	14 206

Prvých 10 univerzít, na ktoré podalo prihlášku najviac študentov

universities.name	**Počet riadkov**
Akadémia médií odborná vysoká škola mediálnej a marketingovej komunikácie v Bratislave	6 132
Technická univerzita vo Zvolene	6 042
Vysoká škola bezpečnostného manažérstva v Košiciach	6 038
Vysoká škola zdravotníctva a sociálnej práce sv. Alžbety v Bratislave	6 035
Univerzita Pavla Jozefa Šafárika v Košiciach	6 019
Bratislavská medzinárodná škola liberálnych štúdií	6 005
Vysoká škola ekonómie a manažmentu verejnej správy v Bratislave	5 965
Katolícka univerzita v Ružomberku	5 900

Vysoká škola múzických umení v Bratislave	5 892
Trnavská univerzita v Trnave	5 883

Študijné odbory, ktoré neúspešne ukončilo najviac uchádzačov

fields_of_study.name	**Počet riadkov**
filozofia	1 612
vnútorné choroby	1 599
liečebná pedagogika	1 581
energetické stroje a zariadenia	1 577
divadelné umenie	1 571
údržba strojov a zariadení	1 557
neurovedy	1 551
ochrana rastlín	1 540
teória vyučovania fyziky	1 528
informatika	1 526

Počet známok, ktoré boli udelené študentom na maturitách

graduation.mark	**Počet riadkov**
1	312 802
2	312 272
3	312 379
4	312 547

Najčastejšie krstné mená

students.name	**Počet riadkov**
Júlia	5 137

Pravoslav	2 734
Bartolomej	2 729
Levoslav	2 717
Fedor	2 710
Miloš	2 708
Bonifác	2 696
Arpád	2 692
Vavrinec	2 689
Mikuláš	2 687

Najčastejšie priezviská

students.surname	**Počet riadkov**
Lubinová	748
Uramová	742
Krajčovičová	727
Kováčiková	725
Novotová	720
Sokolová	715
Sklenárová	714
Húšková	713
Filipová	713
Cesnaková	712

Počet študentov, ktorý majú emailové adresy na jednotlivých doménach

students.email	**Počet riadkov**
szm.sk	91 262
outlook.com	91 173
sme.sk	91 115
zoznam.sk	91 112

gmail.com	91 071
azet.sk	91 040
pokec.sk	90 885
post.sk	90 776
pobox.sk	90 579
inmail.sk	90 548
centrum.sk	90 439

5. Sumarizácia numerických atribútov

Nasledujúce tabuľky zobrazujú sumarizáciu jednotlivých atribútov. Keďže numerický atribút je iba jeden, sú sumarizované aj cudzie kľúče a dátumy v rokoch.

Známky z maturít a dátumy

Názov atribútu	MIN	MAX	AVG	Q1	Q2	Q3
YEAR(students.birth_at)	1916	2016	1965.74	1941	1966	1991
YEAR(graduations_from_ss.graduated_at)	1916	2017	1991.27	1978	1998	2010
graduations_from_ss.mark	1	4	2.4997	1	2	4
YEAR(awards.awarded_at)	1916	2017	1991.20	1978	1998	2010
YEAR(graduations.started_at)	1916	2017	1982.79	1966	1987	2003
YEAR(graduations.finished_at)	1917	2017	1999.66	1992	2005	2013
YEAR(registrations.changed_at)	1916	2017	1991.24	1978	1998	2010

Cudzie kľúče tabuliek, z hodnôt sú uvedné iba celé čísla

Názov atribútu	MIN	MAX	AVG	Q1	Q2	Q3
awards.award_level_id	1	5	3	2	3	4
awards.award_name_id	1	87	44	22	44	66
awards.student_id	3	999 998	500 081	250 153	500 414	750 432
graduations_from_ss.student_id	1	1 000 000	499 777	249 392	499 466	749 743
graduations_from_ss.subject_id	1	21	11	6	11	16
students.secondary_school_id	1	71	35	18	36	54
graduations.student_id	3	1 000 000	499 560	249 230	499 275	749 753

graduations.fos_at_university_id	1	7 000	3 499	1 748	3 498	5 249
registrations.student_id	2	999 995	500 464	250 417	501 859	750 079
registrations.fos_at_university_id	1	7 000	3 503	1 753	3 510	5 258
registrations.status_id	1	6	3	2	3	5
fos_at_universities.university_id	1	35	17	9	18	27
fos_at_universities.field_of_study_id	1	328	165	84	166	247

6. Zhodnotenie

Analytický report zodpovedá rozsahom a podrobnosťou zadaniu.