Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Ilkovičova 2, 842 16 Bratislava 4

Analytický report

Adam Žúrek

Študijný program: Informatika
Ročník: 2
Predmet: Databázové systémy
Cvičenie: Ut 8:00
Cvičiaci: Ing. Matej Kloska
Ak. rok: 2016/2017

1. Opis spôsobu generovania dát

Dáta som generoval pomocou skriptu naprogramovaného mnou v programovacom jazyku Python. Pri generovaní dát som použil užitočnú knižnicu Faker. Vygenerované dáta sa uložili do textového súboru, odkiaľ som ich manuálne skopíroval ako príkazy do PgAdmina. Celý funkčný generátor prikladám k odovzdanej časti projektu.

2. Počty riadkov

Tabuľky som naplnil rôznym počtom riadkov, kde dve som naplnil miliónom riadkov.

1. Classroom …… 50
2. Grade …… 1000000
3. Scholarship …… 107237
4. Scholarship_type …… 5
5. Student …… 1000000
6. Subject …… 11
7. Teacher …… 30
8. Teaching …… 100

3. Sumarizácia tabuliek

3.1. Numerické atribúty

Grade – value

Min: select min(value) from grade; =1

Max: select max(value) from grade; =5

Priemer: select avg(value) from grade; =2.999902

Dolny kvantil: select percentile_cont(0.25) WITHIN GROUP (ORDER BY value) from grade; =2

Median: select percentile_cont(0.5) WITHIN GROUP (ORDER BY value) from grade; =3

Horny kvantil: select percentile_cont(0.75) WITHIN GROUP (ORDER BY value) from grade; =4

Scholarship – amount

Min: select min(amount) from scholarship; =500

Max: select max(amount) from scholarship; =1490

Priemer: select avg(amount) from scholarship; = 995.3657

Dolny kvantil: select percentile_cont(0.25) WITHIN GROUP (ORDER BY amount) from scholarship; =750

Median: select percentile_cont(0.50) WITHIN GROUP (ORDER BY amount) from scholarship; =1000

Horny kvantil: select percentile_cont(0.75) WITHIN GROUP (ORDER BY amount) from scholarship; =1240

Student – grade

Min: select min(grade) from student; =1

Max: select max(grade) from student; =5

Priemer: select avg(grade) from student; =3.000208

Dolny kvantil: select percentile_cont(0.25) WITHIN GROUP (ORDER BY grade) from student; =2

Median: select percentile_cont(0.50) WITHIN GROUP (ORDER BY grade v) from student; =3

Horny kvantil: select percentile_cont(0.75) WITHIN GROUP (ORDER BY grade) from student; =4

Subject – score

Min: select min(score) from subject; =1.2

Max: select max(score) from subject; =5.0

Priemer: select avg(score) from subject; = 2.53636

Dolny kvantil: select percentile_cont(0.25) WITHIN GROUP (ORDER BY score) from subject; =1.95

Median: select percentile_cont(0.50) WITHIN GROUP (ORDER BY score) from subject; =2.3

Horny kvantil: select percentile_cont(0.75) WITHIN GROUP (ORDER BY score) from subject; =2.9

Teaching – weekday

Min: select min(weekday) from teaching; =1

Max: select max(weekday) from teaching; =5

Priemer: select avg(weekday) from teaching; =2.76

Dolny kvantil: select percentile_cont(0.25) WITHIN GROUP (ORDER BY weekday) from teaching; =1

Median: select percentile_cont(0.50) WITHIN GROUP (ORDER BY weekday) from teaching; =3

Horny kvantil: select percentile_cont(0.75) WITHIN GROUP (ORDER BY weekday) from teaching; =4

3.2. Nominálne atribúty

Scholarship -> Scholarship_type -> type

SELECT s.scholarship_type_id, count(s.scholarship_type_id) FROM scholarship s
GROUP BY s.scholarship_type_id
ORDER BY s.scholarship_type_id

Merit-based Scholarship	=21319
Need-based Scholarship	=21596
Athletic Scholarship	=21520
Individual Scholarship	=21541
Scholarship For Ethnic Minorities	=21261

Teaching -> weekday

SELECT weekday, count(weekday) FROM teaching
GROUP BY weekday
ORDER BY weekday

1 (Monday)	=26
2 (Tuesday)	=23
3 (Wednesday)	=16
4 (Thursday)	=19
5 (Friday)	=16

Student -> grade

SELECT grade, count(grade) FROM student
GROUP BY grade
ORDER BY grade

1	=200079
2	=199987
3	=199875

4	=199765
5	=200294

4. Záver

Vytvoril som si tabuľky manuálne podľa SQL príkazov a vygeneroval som ich dáta pomocou môjho scriptu. V dokumente som opísal celý postup môjho generovania dát so stručným opisom a opisom počtu riadkov. Pre každý numerický atribút som uviedol všetky potrebné štatistické prvky a každému nominálnemu atribútu početnosť jednotlivých hodnôt. K tabuľke s miliónom prvkov som uviedol postup SQL, akým som sa k daným číslam dopracoval.