Analytický report

Databázové systémy

Meno: **Tomáš Belluš**

E-mail: **tomas.bellus@gmail.com**

Cvičiaci: **Ing. Ondrej Kachman**

Cvičenie: **štvrtok 13:00**

1 Generovanie dát

Na generovanie dát som využíval viaceré stránky (vypisané v kapitole **Použité generátory,** ďalej budem používať íslo stránky) skombinované s vlastným skriptom v pythone a následné spojenie csv súborou do jedného v programe LibreOffice Calc a na stránke convertcsv.com som využil konvertor CSV to SQL. Z dôvodu obmezdzení v pgadmin3 pri vysokých množstáv pridaných dát rádovo stá tisíce som mohol naraz pridať do tabuľky maximálne 200000 riadkov s 1000 riadkami na jeden INSERT.

Skriptom v pythone som generoval všetky cudzie kľúče a zväčša číselné hodnoty ako ceny typu float. Zložitejšie dáta ako unikátne kódy, dátumy, mená a pod. som generoval prostredníctvom spomínaných stránok.

Všetky csv súbory som kovertoval do sql súboru na stránke 3.

Tabuľka *agents*

Atribúty tabuľky:

Id	Name	Phone	Started_at	Rating

Počet riadkov: ***10 000***

Tabuľku som generoval na stránke 3 a to príkazom *name,phone,pick(1|2|3|4|5)*. Atribút *started_at* typu *date* som generoval na stránke 1 v rozmedzí 2001 – 2015. Musel som použiť externý generátor, keďže stránka 3 neponúka výber rozmedzia dátumov.

Nominálne atribúty:

rating	**počet**
1 = amateur	2023
2 = satisfactory	2003
3 = good	1936
4 = persuasive	2037
5 = experienced	2001

Tabuľka *clients*

Atribúty tabuľky:

Id	Name	Phone	Email

Počet riadkov: ***200 000***

Zákazníkov som generoval prostredníctvom stránky 3 príkazom *name,phone,email.*

Tabuľka *locations*

Atribúty tabuľky:

Id	Street	City

Počet riadkov: ***10 000***

Lokality nehnuteľností som generoval rovnako na stránke 3 a to príkazom *street,pick(<všetky mestá uvedené nižšie>).*

Nominálne atribúty:

City	**počet**
Bratislava	877
Zahorská Bystrica	905
Stupava	917
Rusovce	934
Jarovce	897
Svätý Júr	928
Ivanka pri Dunaji	860
Bernolákovo	914
Marianka	892
Tomášov	888
Pezinok	988

Tabuľka *arrangements*

Atribúty tabuľky:

Id	Balcony	Rooms	Toilets	Floors	Furniture	Pool	Garden

Počet riadkov: ***448***

Na generovanie zostáv som použil stránku, kde som povyplňoval políčka podľa typu pola ako *boolean* a *number* s definovaním minima a maxima. Zostáv je maximálne 448.

Numerické atribúty:

	Rooms	Toilets	Floors
Min()	1	1	1
Max()	7	2	2
Avg()	4	1.5	1.5
Horný_kvartil()	6	2	2

Median()	4	1.5	1.5
Dolný_kvartil()	2	1	1

Tabuľka *estates*

Atribúty tabuľky:

Id	Arrangement_id	Location_id	Name	Status	Category	Land	Build_at	price

Počet riadkov: ***1 000 000***

Nehnuteľnosti som generoval na viac častí. 1000000 atribútov *name* som vygeneroval na stránke 3 a to typu GUID (globally unique identificator). Atribút *build_at* som generoval cez stránku 1 v rozmedzí 2008 – 2016. Atribúty *status* a *category* som generoval príkazom *pick(1|2),pick(1|2|3)* na stránke 3. zvyšné atribúty som generoval vo vlastnom pythonovom skripte. Všetky csv súbory som pospájal do 5 csv súborov po 200000 riadkoch.

Numerické atribúty:

	Arrangement_id	Location_id	Land	Price
Min()	1	1	56	100000.05
Max()	448	10000	400	299999.83
Avg()	224.64	5000.74	224.06	200071.14
Horný_kvartil()	337	7495	314	250172.37
Median()	225	5003	228	200105.45
Dolný_kvartil()	113	2503	142	150092.26

Nominálne atribúty:

Status	**počet**
1 = new	501057
2 = second hand	498943
Category	**počet**
1 = house	333353
2 = flat	333165
3 = villa	333482

Tabuľka *open_houses*

Atribúty tabuľky:

Id	Client_id	Estate_id	Agent_id	Time_at

Počet riadkov: ***150 000***

Prezentácie nehnuteľností som generoval cez vlastný python skript (všetky cudzie kľúče) a atribút *time_at* som generoval na stránke 1 v rozmedzí 2017 – 2019.

Numerické atribúty:

	Client_id	Estate_id	Agent_id
Min()	2	4	1
Max()	200000	999997	10000
Avg()	100066.51	500346.29	4994.24
Horný_kvartil()	150023.25	750587.25	7502
Median()	100121.5	500007.5	4991
Dolný_kvartil()	50242.75	250378	2482

Tabuľka *sold_estates*

Atribúty tabuľky:

Id	Agent_id	Price	Sold_date	Sold_to

Počet riadkov: ***200 000***

Túto tabuľku predaných nehnuteľností som generoval cez moj vlastný python skript (cudzí kľúč a atribút *price*), na stránke 1 (dátumový atribút *sold_date* v rozmedzí 2001 – 2014) a meno kupcu – atribút *sold_to* som generoval na stránke 3 príkzom *name.*

Numerické atribúty:

	Agent_id	Price
Min()	1	100001.33
Max()	10000	299997.4
Avg()	5009.25	199923.16
Horný_kvartil()	7517	249852.45
Median()	5020	199842.59
Dolný_kvartil()	2509	149842.36

2 Použité generátory

1. http://random-date-generator.com/
2. https://www.mockaroo.com/
3. http://www.convertcsv.com/generate-test-data.htm