

SPDB - dokumentacja

Temat

Załadowanie danych udostępnionych w konkursie *Predict which 311 issues are most important to citizens* (<http://www.kaggle.com/c/see-click-predict-fix>) do wybranego SZPBD razem z odpowiednimi danymi przestrzennymi (mapy) i danymi dotyczącymi statystyk obszaru pokrytego przez dane (dane statystyczne dotyczące ludności, służby zdrowia, przychodu, itp.). Dodatkowo należy też przygotować zapytania przestrzenne ukazujące różne aspekty zebranych danych.

Wykorzystane technologie

- Mapa – Google Maps
- Baza danych – MySQL
- Klient – JavaScript, jQuery, Google Maps API
- Serwer – GlassFish, Java EE

Charakterystyka zgromadzonych w bazie danych

Zgromadzono dane z dwóch źródeł:

1. *Predict which 311 issues are most important to citizens* (<http://www.kaggle.com/c/see-click-predict-fix>)
2. School Districts – Elementary (<http://www.census.gov/geo/maps-data/data/gazetteer2014.html>)

Dane 1

Dane dotyczące zgłoszeń problemów, wandalizmów itp. mieszkańców USA na terenie czterech miast: Chicago, New Haven, Oakland i Richmond.

Dane zostały pobrane w formacie csv i zawierały następujące dane:

id - a randomly assigned id

latitude - the latitude of the issue

longitude - the longitude of the issue

summary - a short text title

description - a longer text explanation

num_votes - the number of user-generated votes

num_comments - the number of user-generated comments

num_views - the number of views

source - a categorical variable indicating where the issue was created

created_time - the time the issue originated

tag_type - a categorical variable (assigned automatically) of the type of issue

Dane odpowiadające wszystkim atrybutom zostały załadowane do bazy. Liczba wszystkich rekordów - 223123

Dane 2

Dane dotyczące okręgów szkół podstawowych w 24 stanach USA: Arizona, California, Connecticut, Georgia, Illinois, Kentucky, Maine, Massachusetts, Michigan, Minnesota, Missouri, Montana, New Hampshire, New Jersey, New York, North Dakota, Oklahoma, Oregon, Rhode Island, South Carolina, Tennessee, Texas, Vermont, Virginia, Wisconsin i Wyoming.

Dane zostały pobrane w formacie txt i zawierały następujące dane:

Column	Label	Description
Column 1	USPS	United States Postal Service State Abbreviation
Column 2	GEOID	Geographic Identifier - fully concatenated geographic code (State FIPS and local education agency number)
Column 3	NAME	Name
Column 4	LOGRADE	Lowest Grade Provided
Column 5	HIGRADE	Highest Grade Provided
Column 6	ALAND	Land Area (square meters) - Created for statistical purposes only
Column 7	AWATER	Water Area (square meters) - Created for statistical purposes only
Column 8	ALAND_SQMI	Land Area (square miles) - Created for statistical purposes only
Column 9	AWATER_SQMI	Water Area (square miles) - Created for statistical purposes only
Column 10	INTPTLAT	Latitude (decimal degrees) First character is blank or "-" denoting North or South latitude respectively
Column 11	INTPTLONG	Longitude (decimal degrees) First character is blank or "-" denoting East or West longitude respectively

Do bazy załadowano dane z kolumn 3, 4, 5, 10 i 11. Liczba wszystkich rekordów – 2155.

Logiczny model danych zaimplementowany w SZPBD

Dane 1

Tabelę przechowującą dane 1 stworzono przy pomocy polecenia SQL:

```
CREATE TABLE issue (  
id INT NOT NULL PRIMARY KEY,  
latLng POINT NOT NULL,  
summary VARCHAR(700),  
description VARCHAR(1000),  
num_votes INT,  
num_comments INT,  
num_views INT,  
source VARCHAR(200),  
created_time DATETIME,  
tag_type VARCHAR(100),  
city VARCHAR(100),  
INDEX(num_votes), INDEX(num_comments),  
INDEX(num_views), INDEX(source), INDEX(created_time), INDEX(tag_type),  
INDEX(city),  
SPATIAL INDEX(latLng)  
) ENGINE=MyISAM;
```

Na dane przestrzenne *latLng POINT* założono indeks przestrzenny *SPATIAL INDEX(latLng)*.

W tabeli issue pojawił się dodatkowy atrybut w stosunku do atrybutów udostępnionych przez dane 1.

Wyniknął on z obserwacji - po wczytaniu wszystkich punktów i wyświetleniu ich na mapie zauważono, że dane dotyczą czterech obszarów wokół miast wyraźnie oddalonych od siebie (co widać na rys. na następnej stronie). Sprawdzając w jakich przedziałach długości geograficznych znajdują się te miasta, za pomocą następujących poleceń zaktualizowano wartości dla kolumny *city*:

Chicago

```
Update issue i set i.city = 'Chicago' WHERE Y(i.latLng) > -89.5 and Y(i.latLng) < -82.5;
```

Richmond

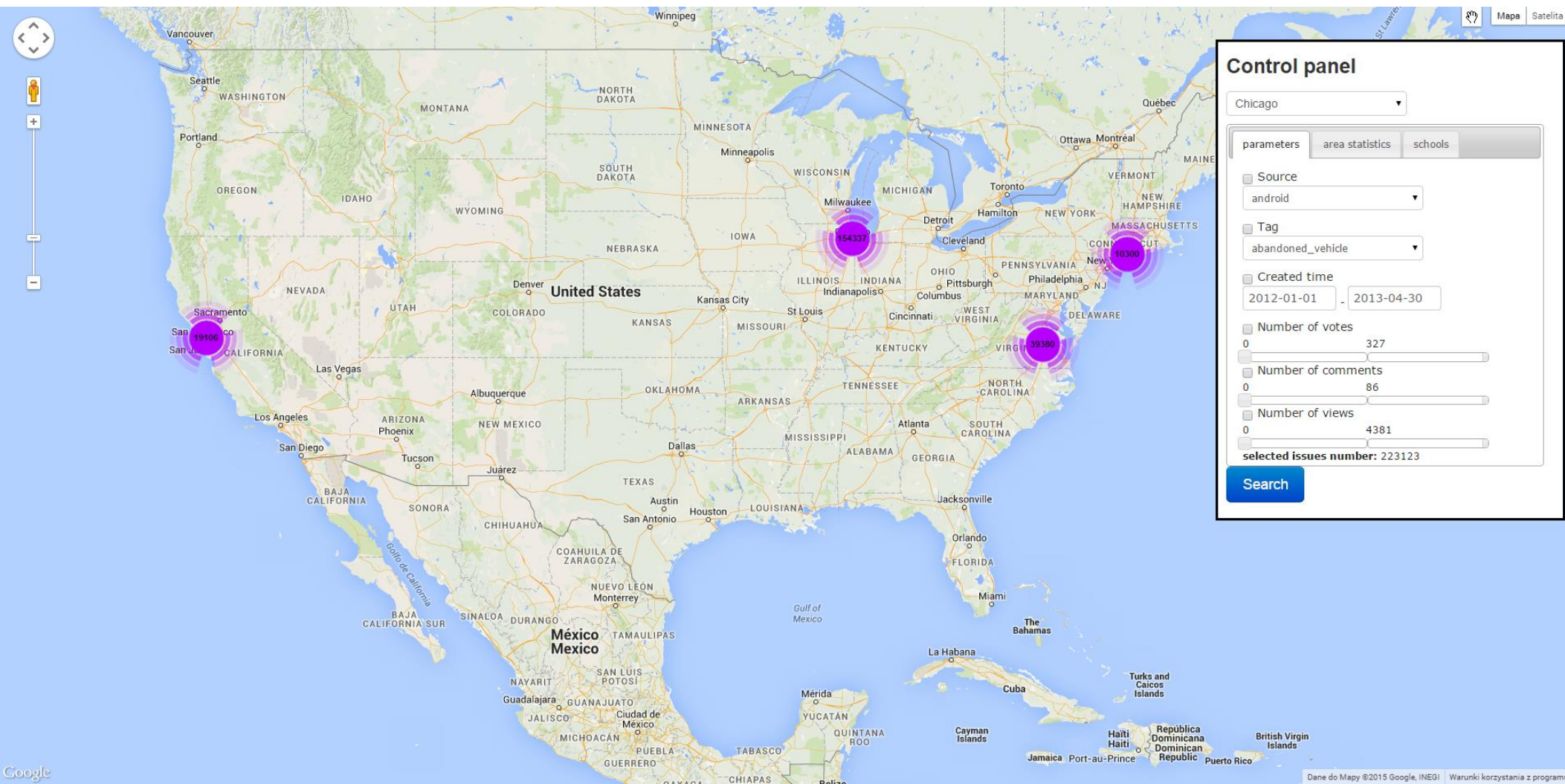
```
update issue i set i.city = 'Richmond' WHERE Y(i.latLng) > -79 and Y(i.latLng) < -75;
```

New Haven

```
update issue i set i.city = 'New Haven' WHERE Y(i.latLng) > -74 and Y(i.latLng) < -68;
```

Oakland

```
update issue i set i.city = 'Oakland' WHERE Y(i.latLng) > -124 and Y(i.latLng) < -119;
```














Dane 2






Tabelę przechowującą dane 2 stworzono przy pomocy polecenia SQL:

```
CREATE TABLE school (  
  id INT NOT NULL PRIMARY KEY,  
  latLng POINT NOT NULL,  
  name varchar(50),  
  lograde VARCHAR(20),  
  higrade int,  
  SPATIAL INDEX(latLng)  
) ENGINE=MyISAM;
```

Na dane przestrzenne *latLng POINT* założono indeks przestrzenny *SPATIAL INDEX(latLng)*.

Diagramy

issue	
 id	int
 latLng	point
 summary	varchar(700)
 description	varchar(1000)
 num_votes	bigint
 num_comments	bigint
 num_views	bigint
 source	varchar(200)
 created_time	datetime
 tag_type	varchar(100)
 city	varchar(100)

school	
 id	int
 latLng	point
 name	varchar(50)
 lograde	varchar(20)
 higrade	int

Zapytania obrazujące różne aspekty zebranych danych

Wyszukiwanie po parametrach

Na podstawie parametrów widocznych na obrazku można wyszukiwać dane. Ze względu na dużą ilość danych (ponad 200 tys.), ograniczono wyszukiwanie do konkretnego miasta. Po wybraniu miasta, określeniu parametrów i wciśnięciu przycisku *Search*, następuje przekierowanie mapy na wybrane miasto, wyświetlenie pinezek odpowiadających danym zgłoszeniom i wyświetlenie liczby zgłoszeń (*selected issues number*).

Po kliknięciu na wybrane miejsce, wyświetla się ramka (*Details*) z dodatkowymi szczegółami. Zdarzają się sytuacje, gdzie występuje kilka zgłoszeń w tym samym miejscu. Wtedy po kliknięciu w to miejsce następuje rozwidlenie pinezek, aby można było wybrać konkretną i wyświetlić jej szczegóły, co pokazano na następnej stronie.

The screenshot displays a Google Map of Chicago. A red pin is placed on a street, and a 'Details' popup window is open. The popup contains the following information:

- tag:** pothole
- description:** State Street between Kinzie and the river was repaved but the mid block crosswalk between Marina Towers and the former IBM Plaza was never repainted. We've been without a marked crosswalk for more than two months now. The signs are still up but no crosswalk markings.
- num votes:** 19
- num comments:** 3
- num views:** 74
- source:** iphone
- created time:** 2012-12-10 01:12:21.0

To the right of the map is a 'Control panel' with the following controls:

- Location:** Chicago (dropdown)
- Tabs:** parameters (selected), area statistics, schools
- Source:** ☒ Source (dropdown: iphone)
- Tag:** ☐ Tag (dropdown: abandoned_vehicle)
- Created time:** ☐ Created time (range: 2012-01-01 - 2013-04-30)
- Number of votes:** ☒ Number of votes (range: 13 - 327)
- Number of comments:** ☐ Number of comments (range: 0 - 86)
- Number of views:** ☐ Number of views (range: 0 - 4381)
- selected issues number:** 2
- Search:** (button)

At the bottom of the page, there is a footer with the following text:

Dane do Mapy ©2015 Google | Warunki korzystania z programu | Zgłoś błąd w mapach

+

-

Mapa

Satelita

Chicago

parameters

area statistics

schools

Source

iphone

Tag

abandoned_vehicle

Created time

2012-01-01

-

2013-04-30

Number of votes

13

327

Number of comments

0

86

Number of views

0

4381

selected issues number: 154337

Search

Google

Dane do Mapy ©2015 Google

Warunki korzystania z programu

Zgłoś błąd w mapach

+

-

Mapa

Satelita

Chicago

parameters

area statistics

schools

Source

iphone

Tag

abandoned_vehicle

Created time

2012-01-01

-

2013-04-30

Number of votes

13

327

Number of comments

0

86

Number of views

0

4381

selected issues number: 154337

Search

Google

Dane do Mapy ©2015 Google

Warunki korzystania z programu

Zgłoś błąd w mapach

W celu przygotowania danych początkowych do obsługi panelu sterującego (dla miast, źródeł, tagów, max. i min. dla *created time*, *number of votes*, *number of comments* i *number of views*) użyto następujących poleceń SQL:

```
select i.source from issue i group by i.source ORDER BY i.source;
```

```
select i.tag_type from issue i group by i.tag_type ORDER BY  
i.tag_type;
```

```
select i.city from issue i group by i.city ORDER BY i.city;
```

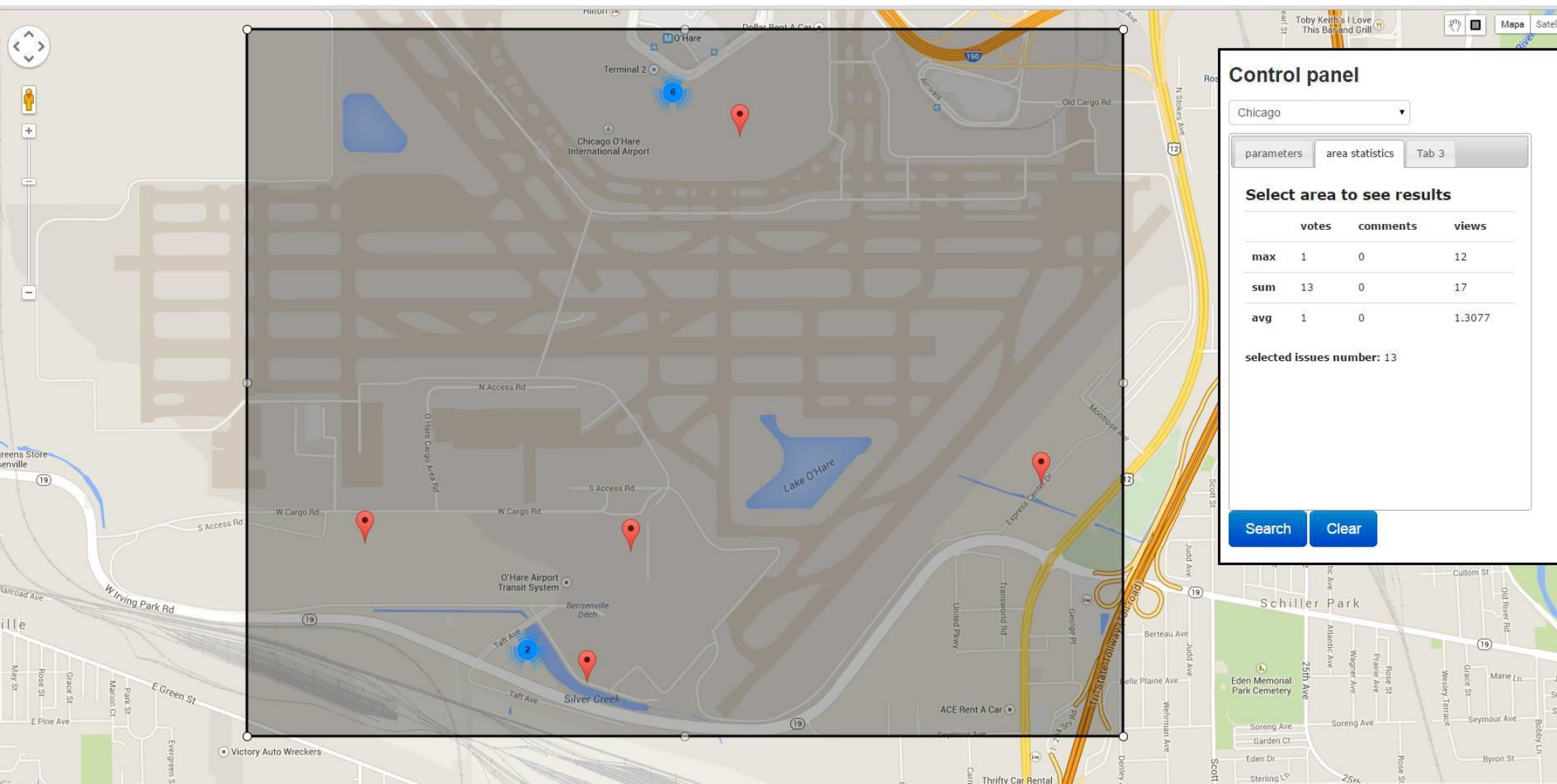
```
select max(i.num_votes) max_votes ,max(num_comments) max_comments,  
       max(i.num_views) max_views, max(i.created_time) max_created_time,  
       min(i.created_time) min_created_time from issue i;
```

Do wyszukania wg parametrów użyto następującego polecenia SQL:

```
SELECT i.id, X(i.latLng) latitude, Y(i.latLng) longitude,  
i.summary, i.description, i.num_votes, i.num_comments,  
i.num_views, i.source, i.created_time, i.tag_type  
FROM issue i WHERE i.city = ?  
AND i.source = ? AND i.tag_type = ?  
AND i.num_votes >= ? AND i.num_votes <= ?  
AND i.num_views >= ? AND i.num_views <= ?  
AND i.num_comments >= ? AND i.num_comments <= ?  
AND i.created_time >= ? AND i.created_time <= ?
```


Wyszukiwanie danych statystycznych po zaznaczonym obszarze

Na podstawie zaznaczonego prostokątnego obszaru, który można przesuwąć oraz którego można zmieniać rozmiar, podawana jest statystyka zgłoszeń występujących w tym obszarze. Statystyka obejmuje: maksymalną liczbę, sumę oraz średnią: głosów, komentarzy i wyświetleń. Dodatkowo podawana jest liczba zgłoszeń objętych obszarem, na podstawie którego obliczana jest statystyka (*selected issues number*).



Do wyszukania statystyki użyto następującego polecenia SQL:

```
select count(*) num_points,  
max(i.num_votes) max_votes, max(i.num_comments) max_comments, max(i.num_views) max_views,  
sum(i.num_votes) sum_votes, sum(i.num_comments) sum_comments, sum(i.num_views) sum_views,  
avg(i.num_votes) avg_votes, avg(i.num_comments) avg_comments, avg(i.num_views) avg_views  
from issue i where st_contains(GeomFromText('Polygon((  
41.586688356972346 -87.84942626953125,  
41.586688356972346 -87.6,  
41.7 -87.6,  
41.7 -87.84942626953125,  
41.586688356972346 -87.84942626953125))'), i.latLng);
```

Współrzędne podawane do tworzenia *Polygon* są parametrami, a w poleceniu podano przykładowy wygląd kompletnego zapytania SQL.

Wyszukiwanie szkół w zaznaczonym promieniu

Na podstawie zaznaczonego okręgiem obszaru wyszukiwane są szkoły występujące w zadanym promieniu. Podawana jest nazwa szkoły, najniższa klasa (przedszkole czy zerówka) oraz ilość klas.

Do wyszukiwania szkół użyto następującego zapytania SQL:

```
select s.name, s.lograde, s.higrade from school s  
WHERE ST_Distance((GeomFromText('POINT(?lat -?lng)'), s.latLng) < ?distance;
```

